DOCUMENT RESUME

ED 050 773 LI 002 835

AUTHOR	Resnikoff, H. L.; Dolby, J. L.
TITLE	Access: A Study of Information Storage and Retrieval with Emphasis on Library Information Systems. Interim Report.
INSTITUTION	R and D Consultants Co., Los Altos, Calif.
SPONS AGENCY	Office of Education (DHEW), Washington, D.C. Bureau of Research.
BUREAU NO	BR-8-0548
PUB DATE	21 May 71
CONTRACT	OEC-0-9-140548-2791(095)
NOTE	225p.
EDRS PRICE	EDRS Price MF-$0.65 HC-$9.87
DESCRIPTORS	*Archives, Books, Indexes (Locaters), *Information Retrieval, *Information Storage, Information Systems, *Library Collections, *Library Materials

ABSTRACT

Chapter I: "Introduction and Summary of Results," stresses the view that the problem of insufficient access is primarily a problem of the great size of the archives to which access is desired. Chapter II: "Levels of Information Storage and Access," is directed toward the problems of library archives and in this context it is access to the content of books and collections of books that is of immediate concern. Chapter III: "Mathematics of Information Distributions," is devoted to the mathematical study of some of the distributions that arise naturally in the study of information systems. Chapter IV: "The Structure of the Back-of-the Book Indexes," is a study of indexes to books in order to determine what structure, if any, they possess. Chapter V: "Algorithmetic Text Indexing," is also exclusively concerned with back-of-the book indexes. Chapter VI: "Amalgamative Access Mechanisms," looks at the problem of discovering possible methods for accessing books. Examples of indexes are appended and the tables included are listed. (NH)

INTERIM REPORT

PROJECT NO. 8-0548

CONTRACT NO. OEC-0-9-140548-2791(095)

ACCESS

A STUDY OF INFORMATION STORAGE AND RETRIEVAL WITH EMPHASIS ON LIBRARY INFORMATION SYSTEMS.

H. L. RESNIKOFF and J. L. DOLBY

R & D CONSULTANTS COMPANY [illegible] 3

Los Altos, California and Houston, Texas

21 May 1971

The research reported herein was performed pursuant to a contract with the Office of Education, U. S. Department of Health, Education, and Welfare. Contractors undertaking such project under Government sponsorship are encouraged to express freely their professional judgement in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent official Office of Education position or policy.

U. S. DEPARTMENT OF

HEALTH, EDUCATION, AND WELFARE

Office of Education
Bureau of Research

ACKNOWLEDGEMENT

The authors wish to express their appreciation to Richard O'Keefe and other members of the library staff of the Fondren Library, Rice University for their generous help and cooperation in the selection of the Fondren Index Sample which provides the central data base of this study. We are also indebted to the late Gerald Mitchell of the Institute for Defense Analyses who aided us in the preparation of the distribution of digraphs; to The Conference Board of the Mathematical Sciences, and particularly the NISIMS Committee, who supported those aspects of this work particularly concerned with accessing mathematical archives; to John W. Tukey, the Statistical Research Techniques Group of Princeton University and the National Science Foundation who supported the work on algorithmic indexing and made available for this study preliminary output from their permuted title listings of the retrospective file of statistical papers; and to M. L. Puri, Department of Mathematics, Indiana University, for his thoughtful contributions to the study of the mathematical models of access systems.

Finally, we should like to acknowledge the contributions of the staff of R & D Consultants Company; William E. Houchin, particularly for his work on the information theoretic aspects of the problem; Val Forsyth for her invaluable contributions to the overall data handling problems; and to Joan Resnikoff and Rena Wells for their painstaking efforts in analysing in fine detail the index structure of the Fondren Index Sample.

TABLE OF CONTENTS

CHAPTER I INTRODUCTION AND SUMMARY OF RESULTS 1

CHAPTER II LEVELS OF INFORMATION STORAGE AND ACCESS 18

CHAPTER III MATHEMATICS OF INFORMATION DISTRIBUTIONS 54

CHAPTER IV STRUCTURE OF BACK OF THE BOOK INDEXES 85

CHAPTER V ALGORITHMIC TEXT INDEXING 112

CHAPTER VI AMALGAMATIVE ACCESS MECHANISMS . . 140

APPENDICES

APPENDIX I ABSTRACT INDEX ENTRIES: A UNIFORM SAMPLE FROM THE FONDREN INDEX SAMPLE

APPENDIX II INDEX PAGE REFERENCE DISTRIBUTIONS FROM THE FONDREN INDEX SAMPLE

APPENDIX III AMALGAMATED ALGORITHMIC INDEX TO ABSTRACTS IN STATISTICS

APPENDIX IV AMALGAMATED ALGORITHMIC INDEX TO ABSTRACTS IN CANCER RESEARCH

LIST OF TABLES

Abstract Index Entries from Computerized Library Catalogs 15

Distribution of Entry Length in Characters (excluding page references) 27

Size in Characters of Various Bibliographic Units 38

Lognormal Standard Deviations 40

Grouped Number of Index Entries for Monographs 75

Distribution of Lexed Words from the *Shorter Oxford Dictionary* by number of vowel strings . 79

Fondren Sample: Fraction of Sample Items Containing an Index, by LC Letter Class . . . 90

Index Access by LC Class 91

Frequency of Index Entries for Items in the Fondren Index Sample 92

Zipf-Mandelbrot Exponent for Index Location Distribution 100

Comparison of High-Frequency Index Entries with LC Subject Headings & Titles 104

Distribution of Index Entries by Word Length Subsample of the Fondren Index Sample 109

Short List of Stop Words arranged by Word Length 117

Excluded Index Terms Referring to One Location 125

Index Entry Length Distribution, *Computerized Library Catalogs* 132

Index Page Location Distribution, *Computerized Library Catalogs* 135

Entry Length Distribution, Algorithmic Index to Statistical Abstracts 150

Abstract Number Location Distribution, Algorithmic Index to Statistical Abstracts . . . 151

Bibliographic Description of the Statistics Index Sample 154

Abstract Entries for the Amalgamated Statistics Index Sample 156

LIST OF FIGURES

Number of Serial Publications with Trend Line 3

Exponential Growth of Scientific Journals and Abstract Journals 5

Index Reference Distributions 10

High Frequency Index References 12

Text and Page Location Distributions 14

Title Length Distribution in Characters, Fondren Index Sample 22

Distribution of Size of Table of Contents for Books in the Fondren Index Sample 23

Index Length Distribution, Fondren Index Sample 25

Distribution of Entry Length in Characters, Statistical Index Sample 28

Distribution of Book Length in Pages for books with Indexes from the Fondren Sample 31

Distribution of Size of University Libraries . 34

The Level Structure of Access Systems 35

Size of Two-Year College Libraries 37

Distribution of Size of Bibliographic Units . 42

Some Access Distributions in Logarithmic Variables 43

Distribution of Number of Characters per LC Subject Heading 46

Distribution of Number of Subject Headings in Fondren Sample (excluding Serials) 46

ALTEXT Macros Ranked by Number of Assembly Language Instructions 49

Word Length Distributions (in Characters) . . 50

Standard Curve of English and Latin Words . . 56

Distribution of Most Frequent Ordered Pairs of English Words 66

Index Entry Distribution from the Fondren Index Sample 67

Page Distribution from the Fondren Sample . . 69

Coefficient of Skewness vs. Coefficient of Kurtosity for Lognormal Functions 74

Shorter Oxford Dictionary 78

Distribution of Income 81

Distribution of Index Length by Number of Index Entries, Fondren Index Sample 94

Number of Index Entries vs. Number of Page References 97

Distribution of Zipf-Mandelbrot Slopes, Subsample of the Fondren Index Sample 101

Distribution of Index Entries by Word Length, Subsample of the Fondren Index Sample 110

Word Frequency vs. Rank, Brown University Standard Corpus of American English 115

Algorithmic Index to Computerized Library Catalogs 121

Index Entry Length Distribution from the Algorithmic Index to Computerized Library Catalogs 134

Index Page Location Distribution from the Index to Computerized Library Catalogs 136

Permuted Title Index Page (Left Side) 144

Permuted Title Index Page (Right Side) 145

Abstract and Abstract Index 146

Cumulative Index to 50 Abstracts (one page) . 147

Entry Length Distribution in Words, Index To Statistical Abstracts 150

Abstract Number Location Distribution, Index to Statistical Abstracts 151

Rank-Frequency of Reference Distribution, Statistical Index Sample 155

Accumulated Index to Books in Statistics (one page) 158

CHAPTER I

INTRODUCTION

AND

SUMMARY OF RESULTS

INTRODUCTION AND

SUMMARY OF RESULTS

This monograph describes work performed by R & D Consultants Company during the first twenty-six months of contract #OEC-0-9-140548-2791(095) with the Office of Education of the Department of Health, Education, and Welfare. The contract is titled "A computer-aided study of access management and collection management in libraries"; its principal objectives are the development of a model for information access and storage systems, and the study of the structure of existing access systems with the intent of augmenting them in significantly useful ways by means of automated processing of machinable data bases.

The concern which underlies this and many other projects is that the rapidly growing body of information stored in library archives is overwhelming the traditional means of obtaining access to it in a reliable, timely, and comprehensive manner.

In fact, general archival collections have been growing in an essentially exponential manner for more than three hundred years, and perhaps for much longer. Figure 1.1, drawn on semilogarithmic graph paper, illustrates this phenomenon for serials noted in the Union List through 1930; this collection has been doubling in size every thirty years. If the trend line is extended back in time, it suggests the publication of a "first" printed serial about 1435, which agrees remarkably well with the invention of printing circa 1440-1456. Although the weight of this evidence is insufficient to convincingly show that serial growth has in fact been exponential since that time, it does support the contention that the exponential growth of archives is a fundamentally long-term property, undoubtedly secondary only to economic and population growth and determined by them. Consequently, it must be anticipated that archival growth will, for the forseeable future, continue to be exponential apart perhaps from fluctuations of minor duration because there does not yet appear to be a significant slackening in either population or long term economic growth for the world as a whole.

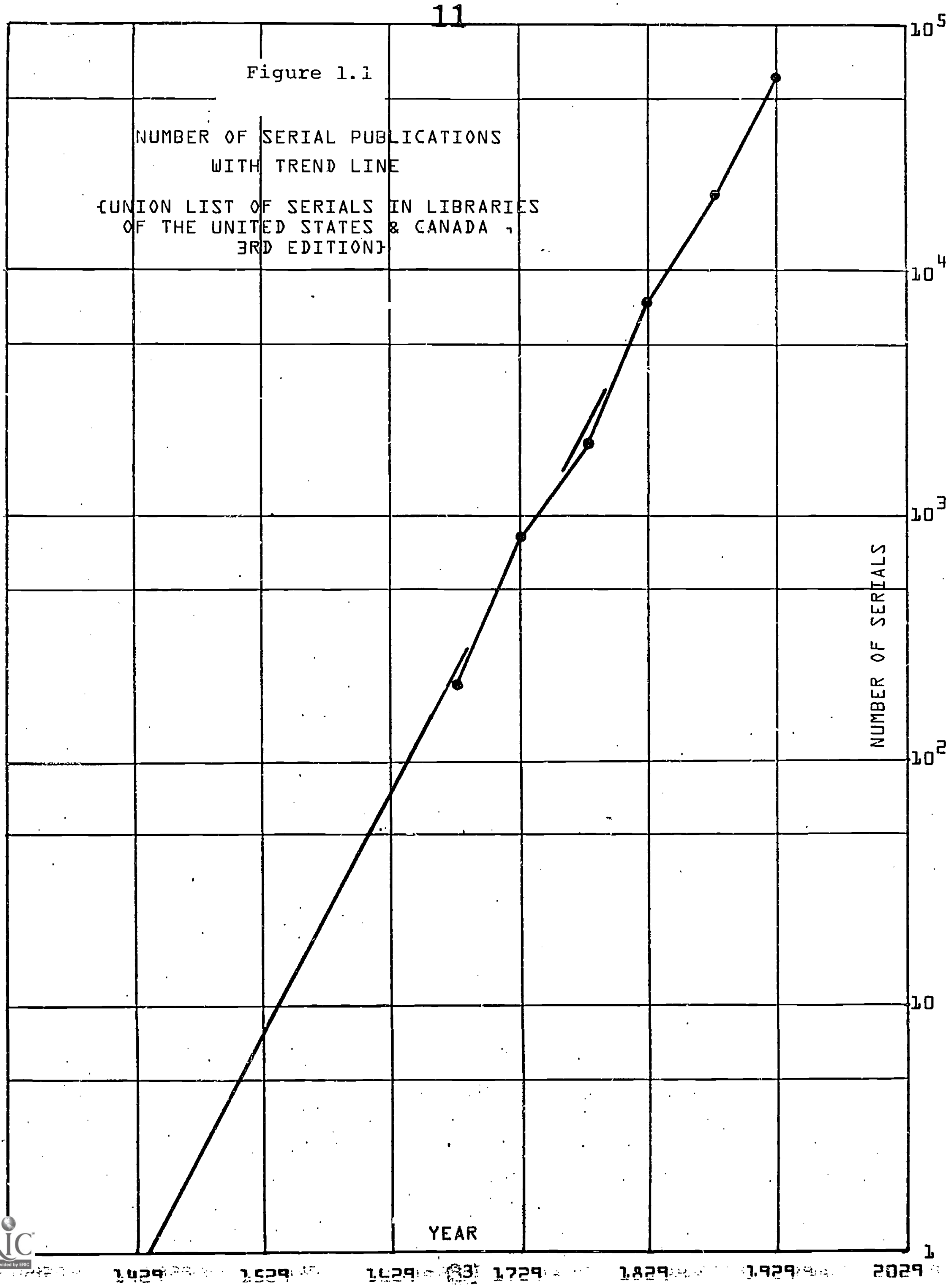


Figure 1.1

NUMBER OF SERIAL PUBLICATIONS
WITH TREND LINE

(UNION LIST OF SERIALS IN LIBRARIES
OF THE UNITED STATES & CANADA ,
3RD EDITION)

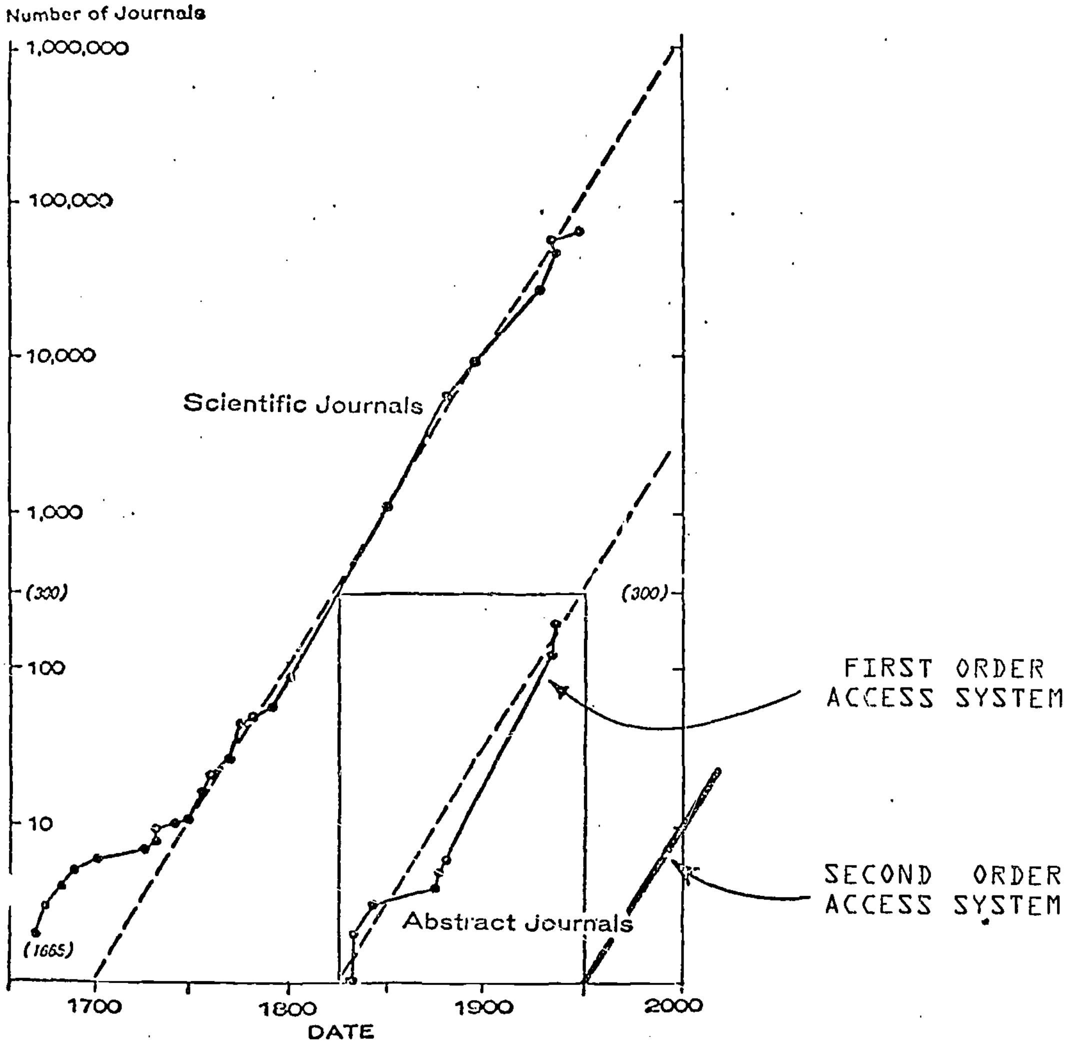
Some insight into why this "information explosion" has received particular emphasis in recent years can be gained from a study of Figure 1.2, adapted from De Solla Price (1) which shows that in addition to the exponential growth of the number of scientific journals since the second half of the seventeenth century, the number of scientific abstract journals has also been growing exponentially, and at the same rate, since their introduction in about 1825. The abstract journals provide access to the larger body of primary scientific journals; the figure shows that the need for this secondary form of p''blication apparently appeared when the number of sci ific journals reached 300. The number of abstract jour ls reached 300 by 1950, making it as difficult to access the abstract journals as it had been to access the primary archive in 1825. This suggests that one of the reasons for the current serious concern about problems of information storage and retrieval is that it is once again necessary to invent an appropriate form of (tertiary) publication which will permit another period of orderly growth of the archive.

If the historical and current trend continues unabated for another fifteen years, there will be about 500,000 different scientific journals in existence, publishing more than 25 million papers each year; similar quantities of information will be spewed forth by other fields of endeavour. It is clearly not the problem of storing this information that makes the prospect of such prolific productivity terrifying; current microform techniques are already sufficient to reduce the physical storage requirements to much less than that presently required to store the current production of journals published in conventional form. Moreover, standardization of microform stores make it possible to implement physical retrieval systems that are faster and cheaper than present typical library storage techniques. Nor is the prospect of having to read all of the published material the significant problem. No scientist since 1800 has had the time to read "all" of the papers published even had he the inclination to do so; the situation is the same in most other fields. The inevitable fact that the fraction of published papers read by an individual is going to drop a few more orders of magnitude is hardly consequential.

The problem posed by the explosion of information is only overwhelming when the difficulty of finding a particular fact or result in the vast sea of information is considered. It is this problem of access to which the work reported here is addressed.

Figure 1.2

Exponential Growth of Scientific Journals and Abstract Journals



Number of Scientific Periodicals (Data from D. J. de Solla Price: *Science since Babylon* {New Haven, 1961], p. 97).

Chapter II introduces a level structured model for access systems, which can be briefly described here. Restricting attention to collections of information expressed in natural languages, size can be reasonably measured by the number of characters, including linguistically necessary interword spaces, contained in the collection. For naturally occurring informational units such as the book title, table of contents, book index, book, and, regarding amalgamated information stores, the university library card catalog and the university library itself, the average size of each informational unit is nearly an integral power of a fixed number K of characters. The value of K is nearly 30. For example, $K \sim 30$ is the average length of a book title measured in characters (as well as the average length of an index entry and of the subject heading information on a Library of Congress catalog card); $K^2 = 874$ characters is approximately the average size of a table of contents; $K^3 = 25,822$ of the average book index, and $K^4 = 763,203$ of the average book. In each case, the average length is remarkably close in value to the power of K in question.

If the size of an information collection is expressed as a power of K, say K^x, then it is convenient to define the <u>level</u> of the collection as the integer closest to x. With this convention, the level of a book title, table of contents, book index, and book is, respectively, 1,2,3, and 4. A university library is of level 8. It therefore appears that the traditional means for retrieving information stored in a book are structured in levels which are equally spaced when measured by their level, that is, when measured by the logarithm of their size.

If one information base is an access system for another, as a book index is for a book, then the <u>order</u> of access is defined to be the difference between their levels. In general, the larger the order, the less expensive is the access system insofar as its construction and maintenance are concerned, relative of course to the cost of obtaining and maintaining the accessed data base; but the smaller the order, the more effective the access system will be in locating specific information and accurately reflecting the content of the accessed archive. For instance, a title list is less expensive and less informative than a collection of abstracts (such as <u>Chemical Abstracts</u>) in specifying the content of journal articles in chemistry; the former is of order 2, the latter of order 1.

The level structure described above will provide a valuable management tool for determining, amongst other things, the reasonable size and cost for a system designed to access a given information base only if the average size of a class of information bases is typical of the distribution of sizes in that class. That this is indeed the case is strongly attested by extensive data sampling studies presented in Chapter II, including the analysis of more than 500,000 index entries occurring in a random sample of books drawn from a medium size university library. All of the evidence reinforces the hypothesis that the distribution of size of information collections belonging to a class is lognormal; that is, the distribution of the logarithm of the size of the informational units belonging to the class is a normal distribution. Each access level corresponds to a different lognormal distribution. It turns out that the variance of the occurring distributions are all nearly the same throughout the entire range from level 1 (titles) to level 8 (university libraries); this means that the distributions depend essentially only on their mean and are therefore characterized by their level. This justifies the use of the notion of level as a measure of an access system.

The principal objective of Chapter III is to show that the lognormal distribution of size of informational units belonging to a class (e.g., titles, books, libraries) is a mathematical consequence of certain reasonable assumptions concerning the "effort" or "cost" of using an item in an access system if the complete system maximizes the output of information per unit effort expended. Our argument is a minor extension of Mandelbrot's derivation of the generalized Zipf-Bradford distribution; cp. Refs. (2), (3). The remainder of the chapter describes general mathematical properties of lognormal distributions with emphasis on the most convenient but nevertheless laborious and not entirely satisfactory technique for fitting lognormal functions to sample data; a number of worked examples which are of independent interest are included.

Unfortunately we do not know a theoretical argument that will produce the equispacing of the level structure of the means of the lognormal distributions associated with an access system; this aspect of the access model rests entirely on observational evidence.

The book is still the most natural informational unit for those concerned with library matters. According to the access model, there are exactly four orders of access associated with collections of information of this size, and, as we have already remarked, there is a traditional access system operating at each of these levels: the index is of order 1, the table of contents of order 2, the title of order 3, and finally, the Library of Congress letter class, which partitions the entire span of written human knowledge into 21 grand categories, is of order 4.

Although these access levels are, in accordance with the prescription of the model, the only ones possible, there are of course many different types of access systems which can function at each of these levels in addition to those just named. For example, a nine page review of a 277 page book provides typical order 1 access for the book of average size. Order 1 access systems most accurately reflect the content of the information collection they access and can moreover form the subsidiary information base from which access systems of higher order (i.e., lower level) can be constructed. Because this procedure obviously cannot be reversed--a low order access system can never be constructed from one of higher order--order 1 access systems deserve special study.

Of the traditional order 1 access systems, the book index is the most amenable to extensive statistical analysis, both because it is found in close proximity to the book text to which it refers (which is generally not the case for book reviews) and because it is naturally composed of a large number of homologous small entities which are suitably arranged for analytical study.

We have investigated three major collections of book indexes. The first contains more than 100 books drawn from the present authors' libraries; although this sample exhibits some variation in subject matter, science and more particularly history, mathematics, and physics are heavily weighted. The second sample consists of 80 current books in statistics and probability theory and comprise what can be thought of as a specialists hand library; it undoubtedly accurately reflects the nature of indexes to books in these fields. All index entries in this sample were committed to machine readable form to permit ready reorganization and analysis of the collection of index entries, of which there were 31,232.

Study of these collections was instrumental in guiding us to the formulation of the access model presented in Chapter II, but their limitations--principally their restriction to few subject areas and the undoubtedly biassed method of their selection, but also their relatively small size--clearly indicated the desirability of carefully selecting a random sample of books and their indexes from a broadly representative archive. The third index sample consists of such a random selection of 706 indexes from the Fondren Library at Rice University. For each book in this sample, copies of the shelf list catalog card, title page, table of contents, and index were made. From this information it is possible to determine the size of the book (in pages, which can then be approximately adjusted to equivalent number of characters) and the precise number of characters in the title, table of contents, and index, which are the three significant traditional book access systems that are normally packaged with the book itself.

Chapter IV describes the structure and properties of traditional back of the book indexes based on a study of these three samples. There are three main conclusions: first, the average number of index entries per index is determined; the result is 836, with relatively little variation throughout the different Library of Congress letter classes. Second, it is shown that the distribution of the number of books as a function of the number of entries in their index is lognormal, providing further support for the access model derived in Chapter II. The remainder of Chapter IV is devoted to a study of the distribution of the number of text references per index term in a given book. The underlying idea is an outgrowth of the simple observation that those index entries that refer to only one text page cannot typify the general content of the book, whereas an entry that refers the reader to 40 or 50 text pages is truly of little specific utility to the reader except insofar as it points out one general topic of the book. It is therefore conceivable that some subset of the index functions as a collection of "key words", specifying the semantic content of the work and serving little further purpose. Were it possible to separate this subset from the other more numerous index entries, the way would be clear for automatic descriptor determination based on a machine readable index; moreover, if the process of constructing the index itself could be automated, iteration of these processes would lead to the descriptors as well and quite possibly to a successful method for man-machine interactive content classification.

Figure 1.3 exhibits the page reference distribution for two books; LB875.C7 was published in 1922, is titled Two Views of Education, and contains 775 index entries, whereas DS423.C85 v4 is the fourth volume of The Cultural Heritage of India, published in the interval 1953-58, and

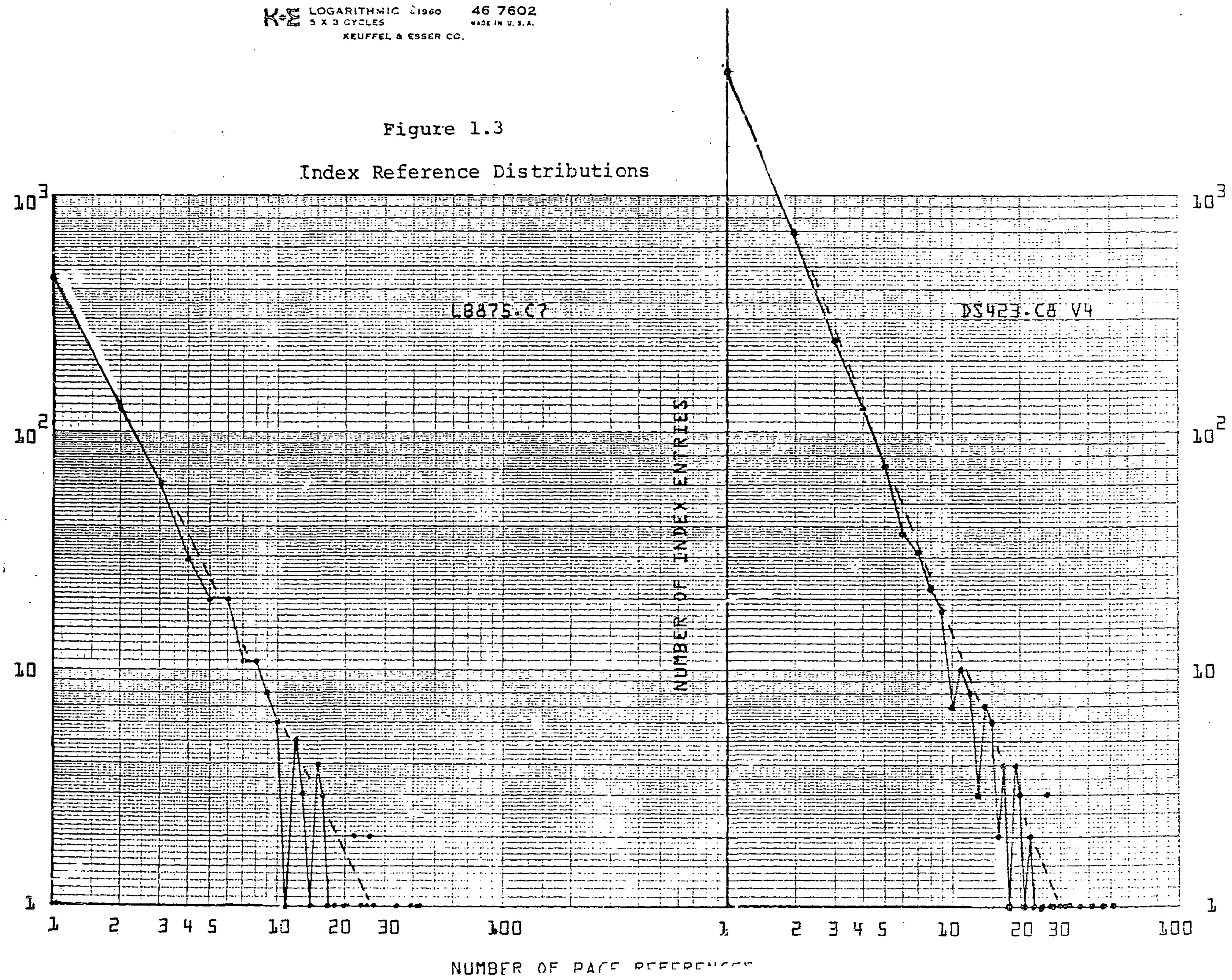


Figure 1.3

Index Reference Distributions

containing 4906 index entries. It is clear from the figure, which is drawn on full logarithmic graph paper, that except for the quite small numbers of entries referring to very many pages both distributions are linear on the graph paper and hence the number of index entries is a power function of the number of page references. From the theoretical considerations in Chapter III one is led to suspect that these graphs ought perhaps to represent lognormal functions, which appear on logarithmic graph paper as parabolas. As we show in the third chapter, the power function, represented by a straight line, is a degenerate form of the lognormal representing parabola. Other book indexes, as for instance that of The 1969 World Almanac, illustrated at the left half of Figure 1.4, do indeed flaunt a tell-tale curvature and can be accurately fitted by a parabola. This is the third major result of Chapter IV: the page reference distribution of index terms is, generally, a lognormal function which may degenerate into a power function.

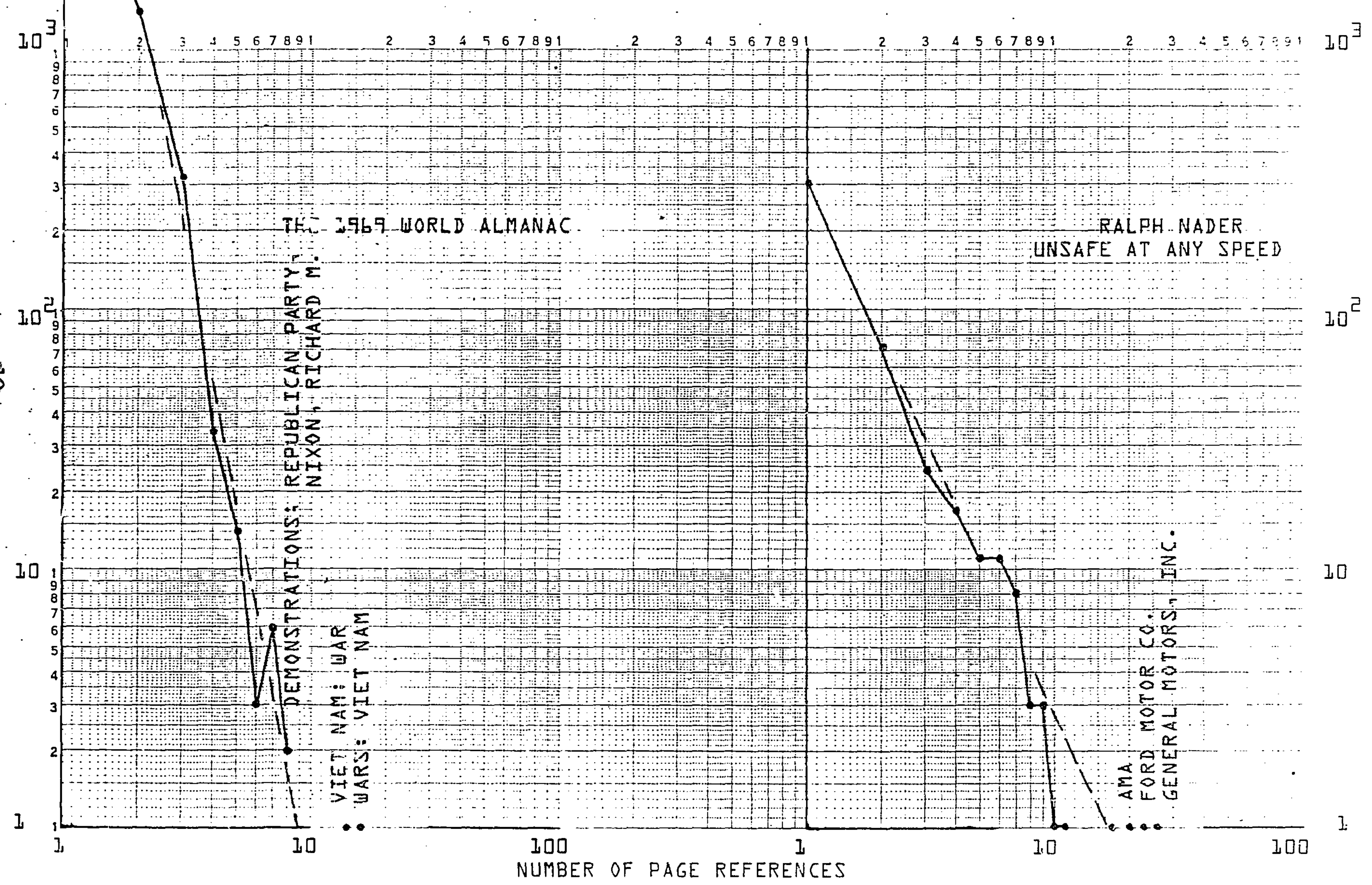
There are about 6600 index entries in The 1969 World Almanac. This book and others like it are more thoroughly and densely indexed than most, but let us for the moment treat this index as an order 1 access system for its text, as usual, but simultaneously consider it as an information base requiring access systems. Then selection of that 1/30 of the index which refers to the largest number of page locations will produce an order 2 access system for the original text, which will be of the size of a table of contents. Repeating this operation leads to the selection of the subset of the index which is about $1/(30)^2 = 1/900$ the size of the index and approximately the size of a title. This process produces about 8 index entries, substantially larger than a title because of the peculiarities of almanacs. Figure 1.4 shows the four most "popular" index entries; in approximately the space of the title they provide an order 3 precis of the book's content which is a not unuseful alternate to the title itself.

The distribution for Nader's Unsafe at any Speed is shown in the same figure; the three most popular index entries again provide a cogently descriptive view of the book's content which is in fact not provided at all by the title.

Chapter IV pursues the study of the effectiveness of the popular index entries as content descriptors through the analysis of a uniform subsample of the Index Sample; for this subsample those index entries which refer to large numbers of text locations have been explicitly listed. The subsample is included as Appendix I.

Figure 1.4

High Frequency Index References



The problem of automatically indexing documents and books has intrigued computer buffs for years. Numerous programs are now available and many learned research papers have been written describing them, and how modest are their demands on the machines that implement them, and how effective they are in satisfying rather imprecisely stated hypothetical requirements of potential users. But as far as we have been able to learn, no commercial or professional publishing house uses machines to index books or papers. The reason that this is so consists of a complex of subreasons not all of which have to do with the adequacy of machine methods, but it is certainly true that the general complexity, inflexibility, and simple inadequacy of these programs have acted as strong deterents to their use. Although the problem is hardly a trivial one, we think that one of the most significant factors hindering the development of indexing algorithms that will rival and surpass human performance is that no one has ever attempted to assess precisely what properties human produced indexes actually have as opposed to what indexers and students of indexing believe ought to be the properties of indexes. The availability of the Fondren Index Sample has made it possible to assess human performance in this area, and to set standards for the performance of machine methods of indexing which are objective, and, insofar as they refer to the structural statistics of indexes rather than their semantic content, also measurable. Elucidation of these common structural characteristics of human indexes have in turn suggested some new approaches to the problem of machine indexing. Chapter V is devoted to one such new method. Figure 1.5 illustrates the text location and page location reference distributions for the algorithmically produced index to Computerized Library Catalogs: Their Growth, Cost and Utility. Based on our study of the Fondren Index Sample we can assert that this algorithmic index is the "right size"; moreover, it is evident from the figure that the reference distributions agree well with typical distributions associated with human indexes. The details of the algorithm as well as of the index referred to by the figure are the subject of the fifth chapter.

Combining these results with those of the previous chapter leads to a new method for obtaining keyword descriptors, which is discussed in the context of the particular algorithmic index exhibited. Indeed, this index consists of 340 entries; an order 1 access system acting on this index should select about 12 entries which would provide an "abstract" of the content of the index. There are 9 entries referring to at least ten page locations, but 14 referring to at least nine. Table 1.1 lists these 14 abstract entries together with the number of pages to which each refers.

Figure 1.5

Text and Page Location Distributions

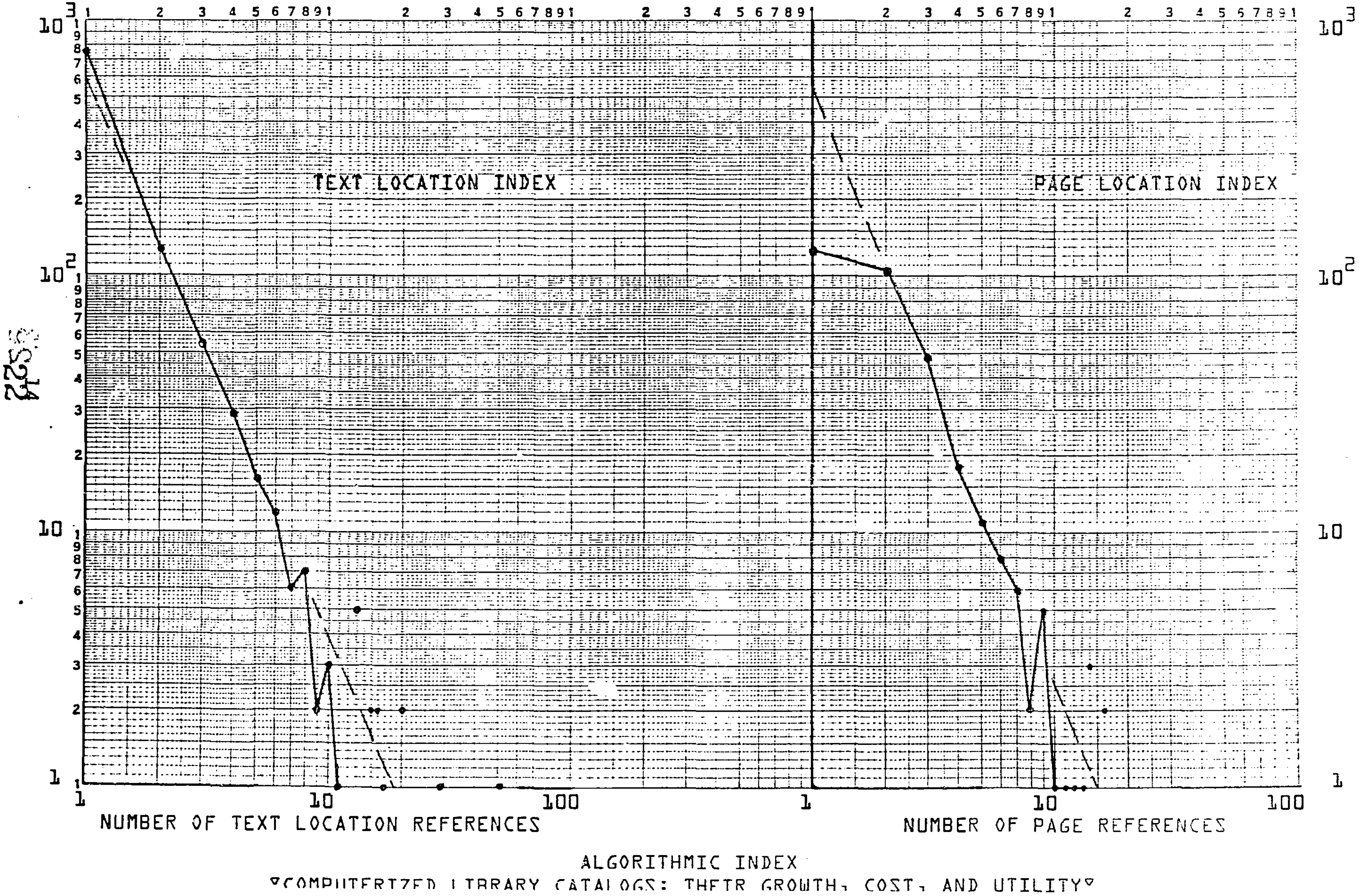


Table 1.1

ABSTRACT INDEX ENTRIES
FROM
"COMPUTERIZED LIBRARY CATALOGS:..."

No. of Page References	Index Entry
16	LC
16	GNP
14	growth rate
14	library catalog
14	machine-readable form
13	Library of Congress
12	gross national product
11	university library
10	exponential growth
9	bibliographic record
9	Fondren, see Rice University
9	Sample, see Fondren Sample, Rice University
9	shelf list
9	Stanford

In this case the abstract entries provide an accurate capsule view of the problems studied in that book as well as a list of the principal sources of information upon which it bases its arguments.

Although there are only four levels available for accessing the text of books and each of these is already served by a traditional access mechanism, there remain many possibilities for repackaging access information in order to serve needs that cannot be met by traditional means. Many of these are amalgamative in the sense that they combine access information associated with numerous comparable unit information stores in a reorganized manner that permits ready selection of the units that are likely to contain specific matter desired by the user. All document information retrieval systems operate in an amalgamative manner, as does the library catalog. Most low order (high level) amalgamative access systems currently in use organize the access information in a sequential fashion based first on date of publication and secondarily according to some scheme of content classification. This is the procedure used to organize professional society abstracting journals (which provide order 1 access); its success depends entirely on the accuracy and excellence of the content classification system and the classifiers who implement it. Such systems, which represent professional consensus concerning significant categorical classifications, are in general partly obsolete, especially in rapidly growing fields such as chemistry where the quantity of published material may double in as few as eleven years. Moreover, although the bulk of the classified material remains stable as the classification system expands and is refined, some fraction of the archival materials, which is likely to include the most innovative work, should be reclassified to account for changes in classification categories and procedures, but due to economic constraints, it never is. This difficulty suggests that it may be desirable to investigate amalgamative access mechanisms which do not depend on external classification structures which are inherently slow to accommodate themselves to change but rather rely on the text terms and systems based on the processing of numerous homologous small items such as index entries, as opposed to text abstracts, which are of special interest because they are less subject to global grammatical constraints and therefore admit a greater variety of potentially useful orderings.

Chapter VI studies two new types of amalgamative access systems. The first consists of the combined indexes to a collection of books, here illustrated by the combined indexes to 80 books in statist'cs and probability theory, already mentioned above in another context. The second is more unusual. We have applied a version of the algorithmic indexing procedure described in Chapter V to two samples of 50 abstracts drawn, respectively, from the Annals of Mathematical Statistics and the Journal of Cancer Research; the results are exhibited and analyzed.

Appendix I displays the order 1 abstract entries (but in some cases involving exceptionally large indexes, only the order 2 abstract entries) from a uniform subsample of the Fondren Index Sample.

Appendix II displays the distribution of the number of index entries as a function of the number of distinct text pages to which each entry refers for the same subsample of the Fondren Index Sample used for Appendix I. These distributions confirm the assertion that the distribution is essentially a power function.

Appendices III and IV are automatically constructed amalgamative indexes to 50 abstracts of papers in statistics and in cancer research respectively. The algorithm and all internal dictionary-like stores used by it is the same for both data sources.

REFERENCES

1. De Solla Price, D. J., Science Since Babylon, Yale University Press, New Haven, 1961.

2. Mandelbrot, B., "An Informational Theory of the Statistical Structure of Language", Communication Theory, Butterworths, London, 1953.

3. Mandelbrot, B., "On the Theory or Word Frequencies and on Related Markovian Models of Discourse," in "Structure of Language and Its Mathematical Aspects", Proc. Symp. Appl. Math. 12(1961).

CHAPTER II

LEVELS OF INFORMATION STORAGE AND ACCESS

LEVELS OF
INFORMATION STORAGE AND ACCESS

In the previous chapter we have stressed the view that the problem of insufficient access is primarily a problem of the great size of the archive to which access is desired. This study is directed toward problems of library archives and in this context it is access to the content of books and collections of books that is of immediate concern although libraries are increasingly becoming archival depositories of other types of information bearing records.

There are technical reasons that make it desirable to restrict attention--at least in a preliminary study such as the present one--to the monograph collection; we will have some useful remarks to make about serials and can also exhibit data supporting the extension of the model that will be proposed to describe the serial collection.

The book is a natural halfway house in the hierarchy of means for storing written information in libraries. Within the book are usually to be found certain standard apparatus which aid in directing the user to the internal location of information with which the book is concerned; these include, in descending order of size, the index, the table of contents, and the title. The library itself is of course a collection of books but it too contains certain apparatus for directing the user to those amongst the many books held that contain information concerning some particular matter; these include, in increasing order of size, the classification system, the reference section, and the card catalog. There are also other types of traditional access means that aid in locating books which contain certain information, including special bibliographies and, too often overlooked, the reference librarians. If indeed size is the predominant factor determining the need for access, then a study of the size of the various natural bibliographic units named above may shed light on the structure, if any, of the traditional access systems and thereby also provide guidelines for those who study the possible ways for increasing and automating the means of access.

We will proceed up the scale of size of the naturally occurring access means associated with books and collections of books, with the intent of determining the statistical distribution of size of each such system; this information will lead in a natural way to the level structured model of access systems briefly described in Chapter I.

Initially limiting our attention to the book itself, there are four systems of interest:

1. Title
2. Table of Contents
3. Index
4. Book Text

In each case we wish to know the mean (average) size of the item in question, measured, let us say, by the number of characters (including the interword space) contained in the item. Moreover, it will turn out to be important to know the distribution of size for each case so that it will be possible to say to what the extent the mean is characteristic of the distribution and also because the distributions will turn out to have an intrinsic connection with the access problem via the intervention of the mathematical discipline known as information theory; this latter aspect of our study will be described in Chapter III.

It is not easy to obtain reliable statistics about the size of bibliographic units; it is especially difficult if general samples that are not restricted to one or a few fields of interest are desired. We have based our book studies on the Fondren Sample, a random sample of 1926 cards drawn from the shelf list catalog of the Fondren Library at Rice University in 1968; it has been described in some detail in Ref. (1). Associated with each shelf list card is one or more monographs; these monographs constitute the sample on which our study is based. It is appropriate to refer to it as a random sample of books from a medium sized university library.

Because we are interested in studying the interaction of the various traditional access systems used in books we have extracted from the Fondren Sample all those books that contain an index (here and throughout all that follows, index will of course mean back of the book index), thus yielding what we have called the Fondren Index Sample, which may reasonably be called a random sample of indexes. There are of course certain unavoidable biasses present in this index sample: the Fondren Library does not have an adequate collection in medicine or law, for instance; it has an exceptionally fine collection in other areas. But, to the best of our knowledge, these samples are the closest in existence to truly random samples of books and of books with indexes belonging to the complete population of all books ever published.

With these preliminaries in mind we can now turn to study the structure of book titles. Figure 2.1 displays the distribution of the number of characters per book title for books from the Fondren Index Sample drawn on lognormal probability graph paper. The mean number of characters per title is 28.15.

Next consider the size of a table of contents measured by the number of characters it contains.

Although the "structure" of a book title is relatively standardized, the same cannot be said of the table of contents. Some books include phrases such as "Chapter 1", others simply record "1" to designate the first chapter, and others do not bother to indicate the chapter ordinal at all. There are tables of contents which include, in addition to a chapter title, relatively extensive descriptions of the text content of a narrative nature; others include section titles. Despite the rather excessive degree of variation that does occur, there are certain components of a table of contents which appear to be nearly invariable in their presence, including the chapter titles and page number designating the beginning of each chapter. We have chosen to define the table of contents as that portion of the material contained in what is normally termed the table of contents that corresponds to the chapter title, excluding from consideration all headings, chapter ordinals, appendices, tables of figures, etc., and page number referents to the location of chapter initial pages. With this convention, a random subsample of 161 tables of contents was selected from the Fondren Index Sample and the number of characters (including interword space characters) was counted for each selected table of contents. It turns out that the mean size of a table of contents defined in this way is 505 characters. Figure 2.2 displays the distribution of table of contents size for this subsample.

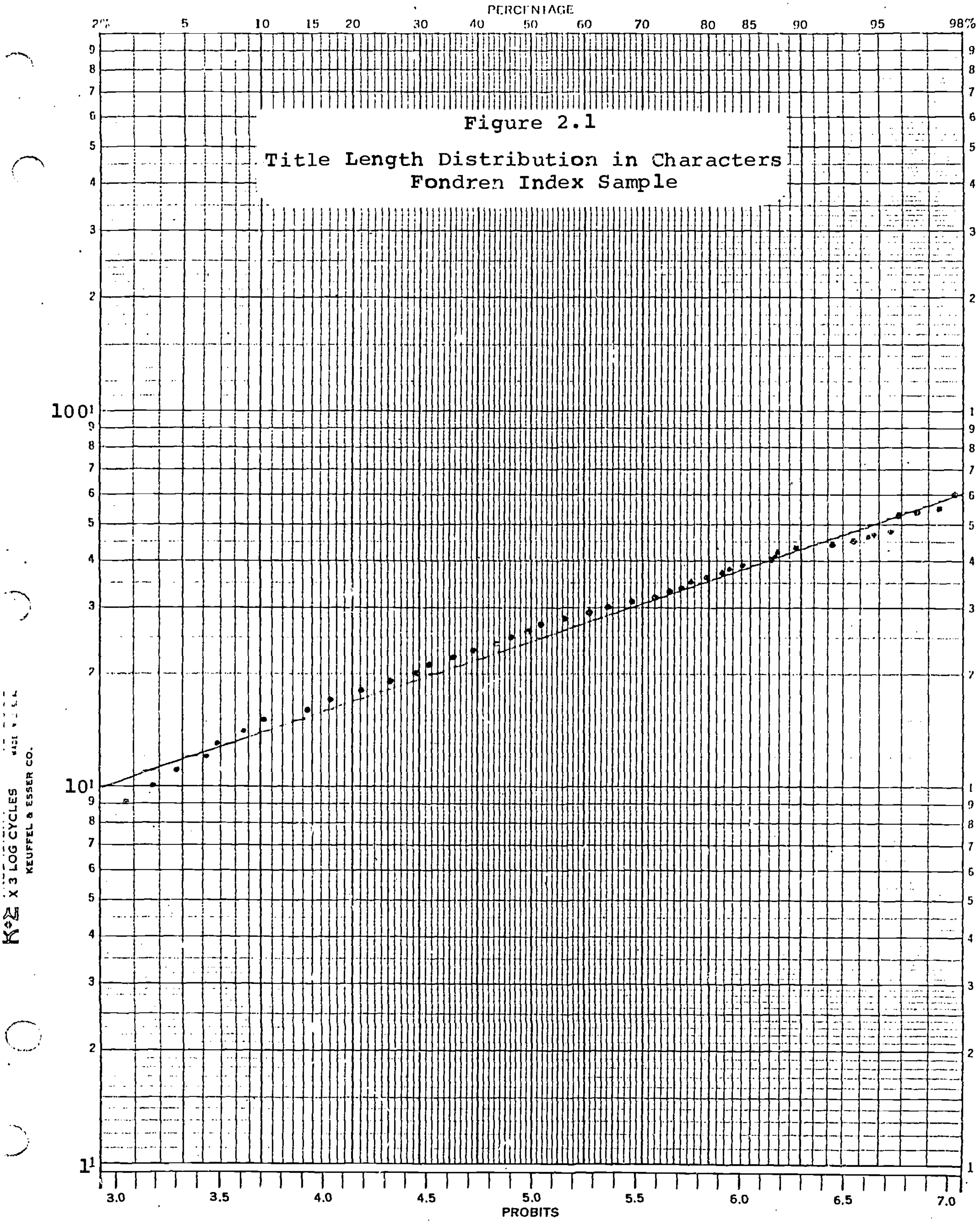


Figure 2.1

Title Length Distribution in Characters
Fondren Index Sample

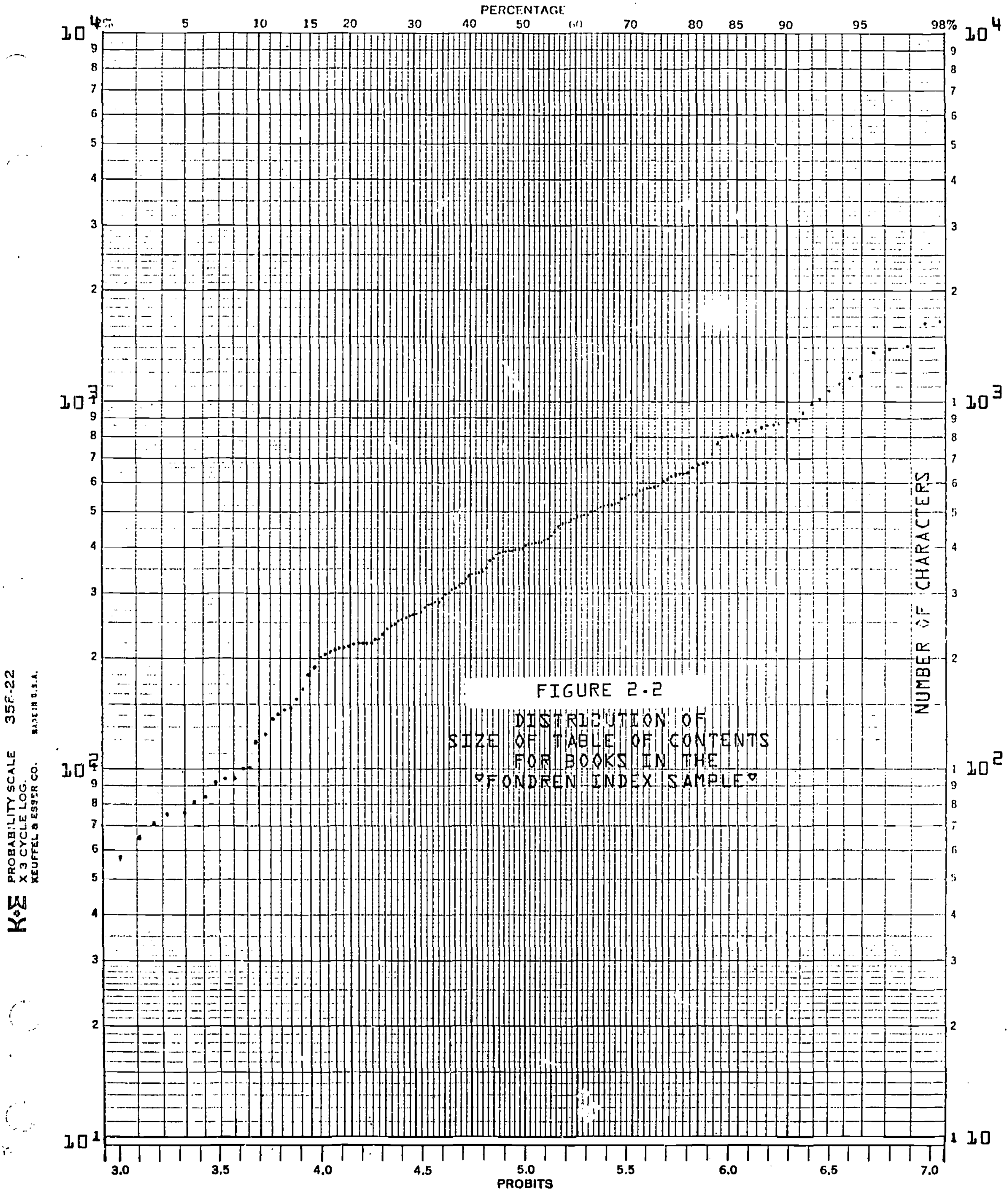


FIGURE 2.2

DISTRIBUTION OF SIZE OF TABLE OF CONTENTS FOR BOOKS IN THE "FONDREN INDEX SAMPLE"

The reader can hardly help but notice that the data exhibited in each of Figures 2.1 and 2.2 fall nearly along a line, and moreover that the two lines have similar slope. The graph paper is so designed that straight lines indicate that the data are drawn from a lognormal distribution, whose properties will be discussed later on in this chapter and extensively in Chapter III; it suffices here to stress that thus far the data indicates that the two lowest levels of distribution of size of book access systems belong to some well known family of statistical distributions and indeed to the same family. We will want to look for this possibility when examining data referring to other access systems.

The index is the next largest access tool traditionally found in books, and from many points of view it is the most important and responsive to the detailed demands of the user. It therefore deserves extensive examination.

The Fondren Index Sample consists of 706 indexes. Chapter IV investigates the relationship of indexed books to the unindexed books in the Fondren Sample and studies such properties of the indexed books as their distribution among the Library of Congress classification categories. Here we are only interested in considerations of size. The mean number of index entries per index is 836.

Figure 2.3 contains the distribution of the number of index entries per book, again on lognormal probability graph paper. It is evident that the data can be accurately approximated by a line and furthermore that the line has a slope which once again is similar to the slope of the lines occurring in the previous two figures. One word of caution: here only the number of index entries is exhibited. Ideally one would wish to measure the size of an index by the number of characters it contains, but it would not be feasible to count the characters in more than half a million index entries. Furthermore, once again the question of which characters to count can not be resolved in a completely unambiguous way. For instance, it is easy to agree whether page reference numbers should be counted, and what to do about consecutive spaces used as separators, but format problems related to multiple entries grouped under a common initial phrase, and inverted order entries demand operational decisions that are not often guided by a clear cut purpose. These problems exist when entries alone are counted, but they are magnified when characters are counted. We have agreed, when counting entries, to count

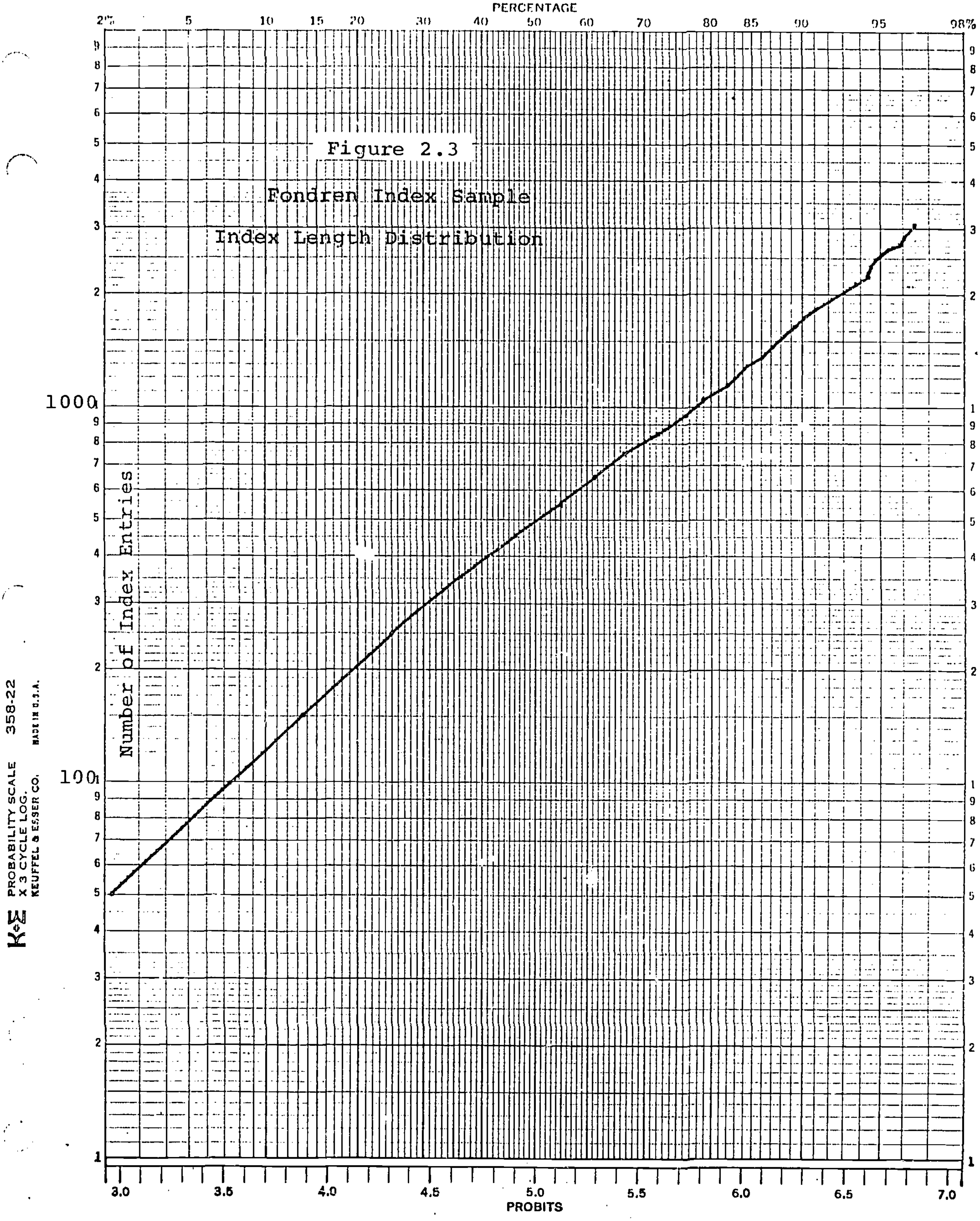


Figure 2.3

Fondren Index Sample

Index Length Distribution

each group of page reference numbers: this defines the index entries, at least as far as their cardinal number is concerned, and provides a relatively clear cut procedure requiring a minimum amount of subjective decision by the persons performing the counting. In order to obtain an approximation to the number of characters contained in an index, a rather indirect procedure was used. We have in a convenient form all of the index entries contained in 80 books in the field of statistics, all printed in a fixed typefont whose characters are of constant width, and printed a fixed number of lines to the page. These characteristics make it possible to count the number of characters in an entry by measuring the length of the entry. This was done for a uniform subsample (comprising about 1.75% of the total Statistical Index Sample of 31,232 index entries). Table 2.1 lists the number of entries consisting of from 1 to 76 characters, and, opposite 77 characters, the number of entries that had at least 77 characters. The mean number of characters per entry, exclusive of page reference numbers but inclusive of interword spaces, is 25.47. Figure 2.4 displays the distribution of size of the entries in the Statistical Index Sample. If we assume that the distribution of size of index entries is independent of the distribution of the number of entries per index, then the average number of characters per index will be the product of the average number of entries by the average number of characters per entry. Using the number for the Statistical Index Sample for the latter, we find that the average number of characters per index (exclusive of page references) is 836 x 25.47 = 21,293. If it be assumed that there are typically three digits and an interword space required to provide the page reference location information, then augmenting the average number of characters per entry by 4 leads to 24,560 characters per index (inclusive of page reference approximation).

The distribution of index entry length for the Statistical Index Sample is, again, lognormal to a high degree of approximation.

Table 2.1

STATISTICAL INDEX SAMPLE

Distribution of Entry Length in Characters
(excluding page references)

No. of Char.	No. of Entries	No. of Char.	No. of Entries
1	0	41	3
2	0	42	3
3	4	43	6
4	2	44	3
5	4	45	5
6	7	46	5
7	8	47	3
8	9	48	4
9	20	49	4
10	22	50	2
11	19	51	4
12	18	52	4
13	14	53	3
14	23	54	6
15	11	55	4
16	28	56	3
17	16	57	2
18	8	--	
19	14	62	1
20	12	63	1
21	17	64	1
22	17	65	1
23	19	--	
24	19	67	1
25	14	--	
26	8	69	1
27	11	--	
28	9	72	1
29	11	73	1
30	7	--	
31	10	75	1
32	9	76	2
33	10	≥77	9
34	6		
35	5		
36	4		
37	5		
38	10		
39	10		
40	9		

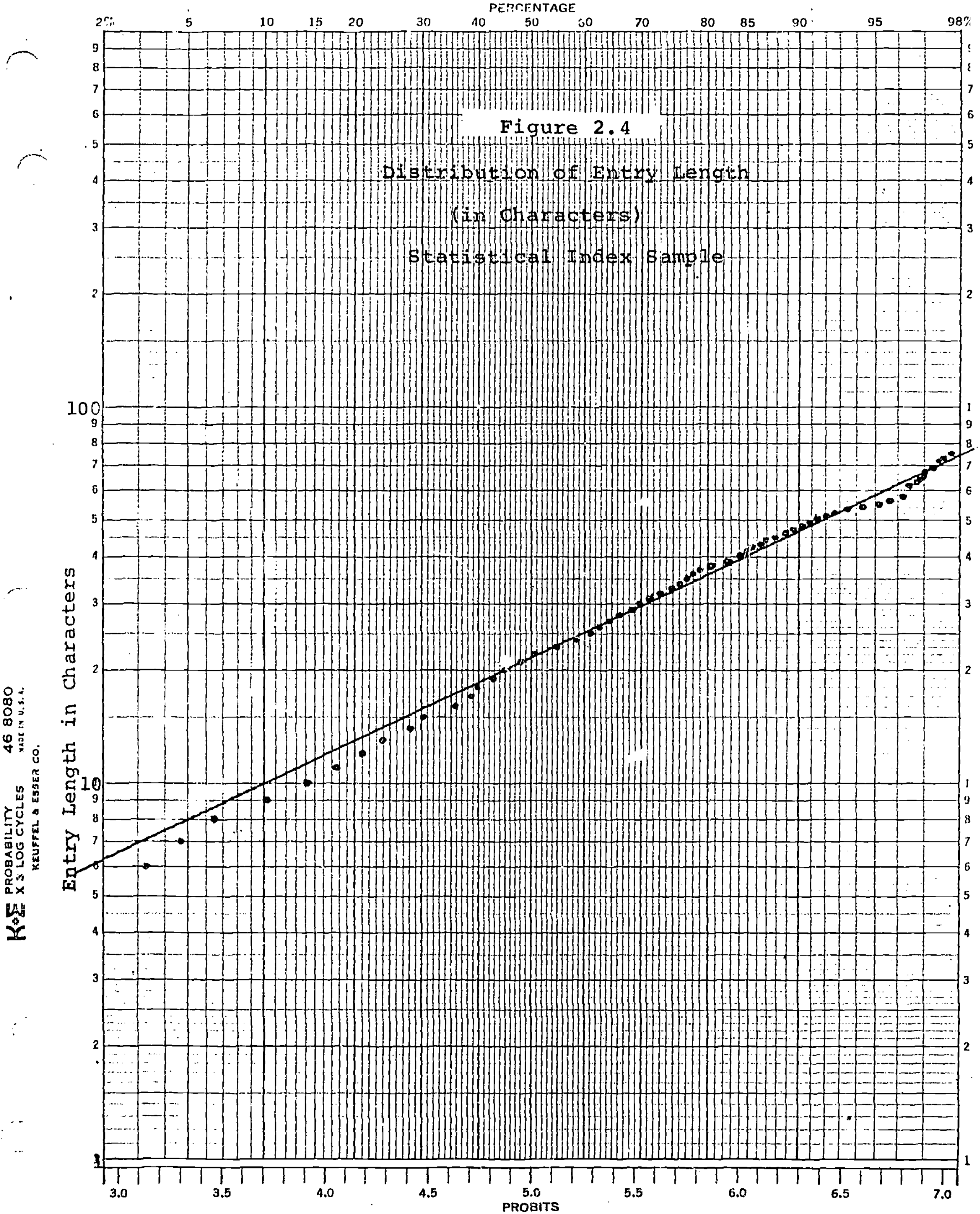


Figure 2.4

Distribution of Entry Length

(in Characters)

Statistical Index Sample

The last of the four natural access tools for monographs is the monograph text itself. It will be even more difficult to estimate the size of a book measured by the number of characters it contains because of the variability of type font and page layout supplemented by the presence of tabular and figured material. Although numerous different and justifiable procedures of making such a size estimate are conceivable, we have once again attempted to choose a method that would be simple and insensitive to subjective judgements of the personnel performing the task in order to improve accuracy but more importantly to make it possible for other workers to reproduce (at least nearly) our results. Regarding book text, there are several levels of analysis that require an increasing amount of extraneous and unstandardized information. The simplest measure, and one that it easily reproduced, is simply to transcribe the arabic number shown on the catalog card designating the number of non-front matter pages. It is difficult to say precisely which pages are represented by that number in each case, but it is unnecessary to do so; we simple agree that this number defines the length of the book in pages. The distribution of book length measured in pages was determined in Ref. (1) for the complete Fondren Sample. The mean number of pages per book is 276.6; the distribution of pages is however not lognormal as is readily seen in Figure 2.5. If the corresponding distribution is plotted just for those books that do have indexes (i.e., for the Fondren Index Sample), the graph in Figure 2.6 results, which shows that the distribution of size of these books is lognormal. This suggests that there may be some intrinsic structural difference between books which contain an index and those that do not. If attention is restricted to the Fondren Index Sample, it turns out that the mean number of pages per book is significantly greater, namely 341.5. The next step in determining the number of characters per book is to find the number of lines per page and their length; this has been studied by Dolby and Jones (Ref.(2)), who found 38 lines of 24 picas as the mean. The final step in obtaining an estimate of book size in characters is to approximate the number of characters per 24 pica line of print; we have analyzed a sample of printed matter and find 63 characters per 24 pica line as the mean. These estimates together imply that an average page of printed text contains 2394 characters, including interword and end of line spaces. Hereafter it will be assumed that there are 2400 characters per page. We have no idea what the effect of tabular and figured material as well as other formatting conventions is on

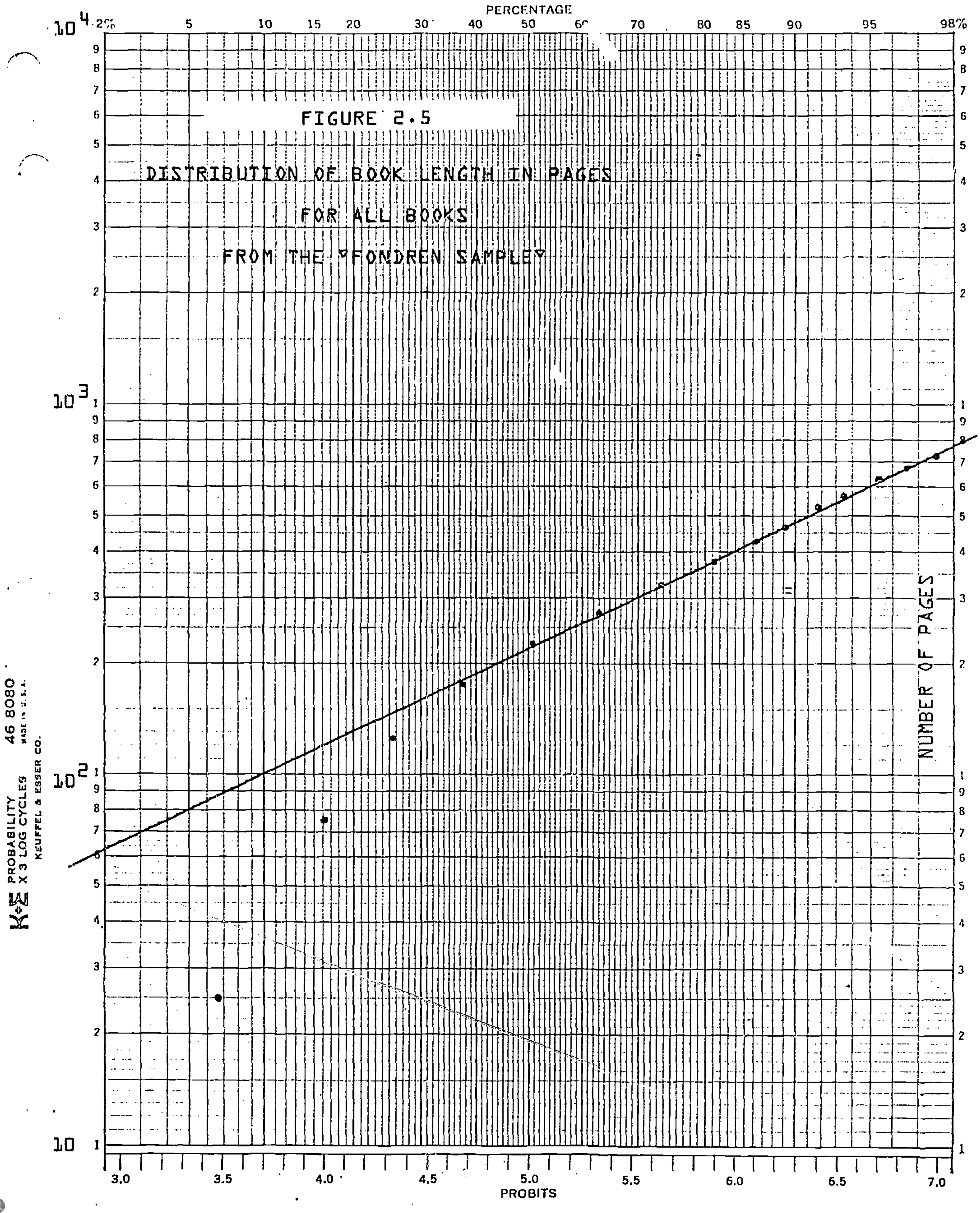


FIGURE 2.5

DISTRIBUTION OF BOOK LENGTH IN PAGES
FOR ALL BOOKS
FROM THE "FONDREN SAMPLE"

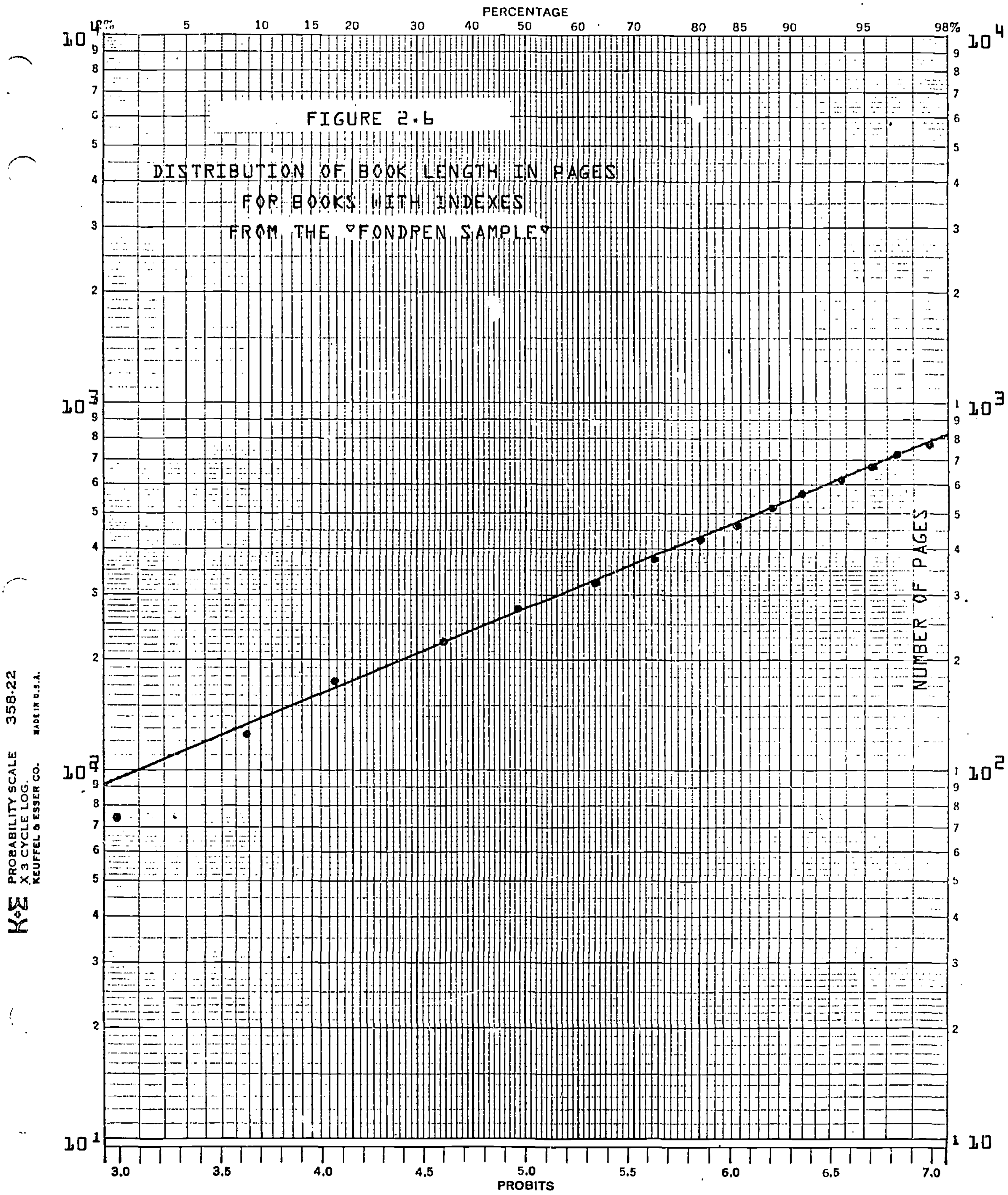


FIGURE 2.6

DISTRIBUTION OF BOOK LENGTH IN PAGES FOR BOOKS WITH INDEXES FROM THE "FONDREN SAMPLE" the estimate of book length in characters; nevertheless, excluding these matters from consideration, we find that the average book in the Fondren Index Sample is 341.5 x 2400 = 819,600 characters in size.

Turning now to collections of books, let us first consider the university library. Here it is essential that the notion "university" be specified in some way so as to enable one to distinguish university libraries from libraries of colleges in a manner consistent with that used for other purposes by governmental agencies and the educational institutions themselves. We implicitly use the definition used by the Office of Education of the Department of Health, Education and Welfare because we use their statistical data book Reference (3) as our source of information about the holdings of college and university libraries.

Unfortunately the data presented in Reference (3) is incomplete; notable ommissions are the University of Chicago and Yale University. Although these ommissions undoubtedly will have some influence on the statistical parameters of the distributions of interest to us, these will most likely be quite minor and in no event can they be expected to change the form of the distribution nor substantially affect its mean or variance.

There is one other defect of the data presented in Reference (3) which is more critical for our concerns. Most state university systems have had their statistics amalgamated; thus it is impossible to determine (from this source) the size of the library of the University of California at Berkeley--only the total number of volumes held in the entire California university system is presented. This unfortunate state of affairs holds for most of the other state systems also and tends both to depress the number of distinct university libraries and inflate the size of those that remain. Two factors permit us to extract useful information from this tabulation despite its amalgamated nature: first, it is easy to obtain lists of all units belonging to a state system (and also for the few private systems that operate more than one campus) and thereby estimate the total number of libraries whose structure must be studied. Second, within state systems there is usually one 'giant' library and a number of much smaller ones; this has the consequence that the departure of the distribution from lognormality, as is shown in Figure 2.7 which we will shortly consider, is diminished when the separate system units are accounted for, and, in view of the smallness of the possible effect, it is not necessary for us to study this difference in detail. Furthermore, we can easily obtain the mean size from the revised estimate of the number of libraries. By adjusting the number of libraries represented in Reference (3) through deletion of the special dental and medical school branches and addition of all general campuses, a total of 201 university libraries is attained. The total number of volumes held in these institutions is 152,230,163 (nearly one for every inhabitant of the United States, and nearly as many as are held by all public libraries), so the mean number of volumes per university library is 757,364. The range in size may appear remarkable to the reader, ranging as it does from some 100,000 volumes to more than 8 million. Figure 2.7 exhibits the size distribution, which, as we have by now come to expect, is lognormal.

Knowing that the average book contains 819,600 characters and assuming that the distribution of book size is independent of the distribution of university library size, we readily find that there are some 620,735,534,400 = 6.2×10^{11}, or approximately 620 billion characters stored in the average university library.

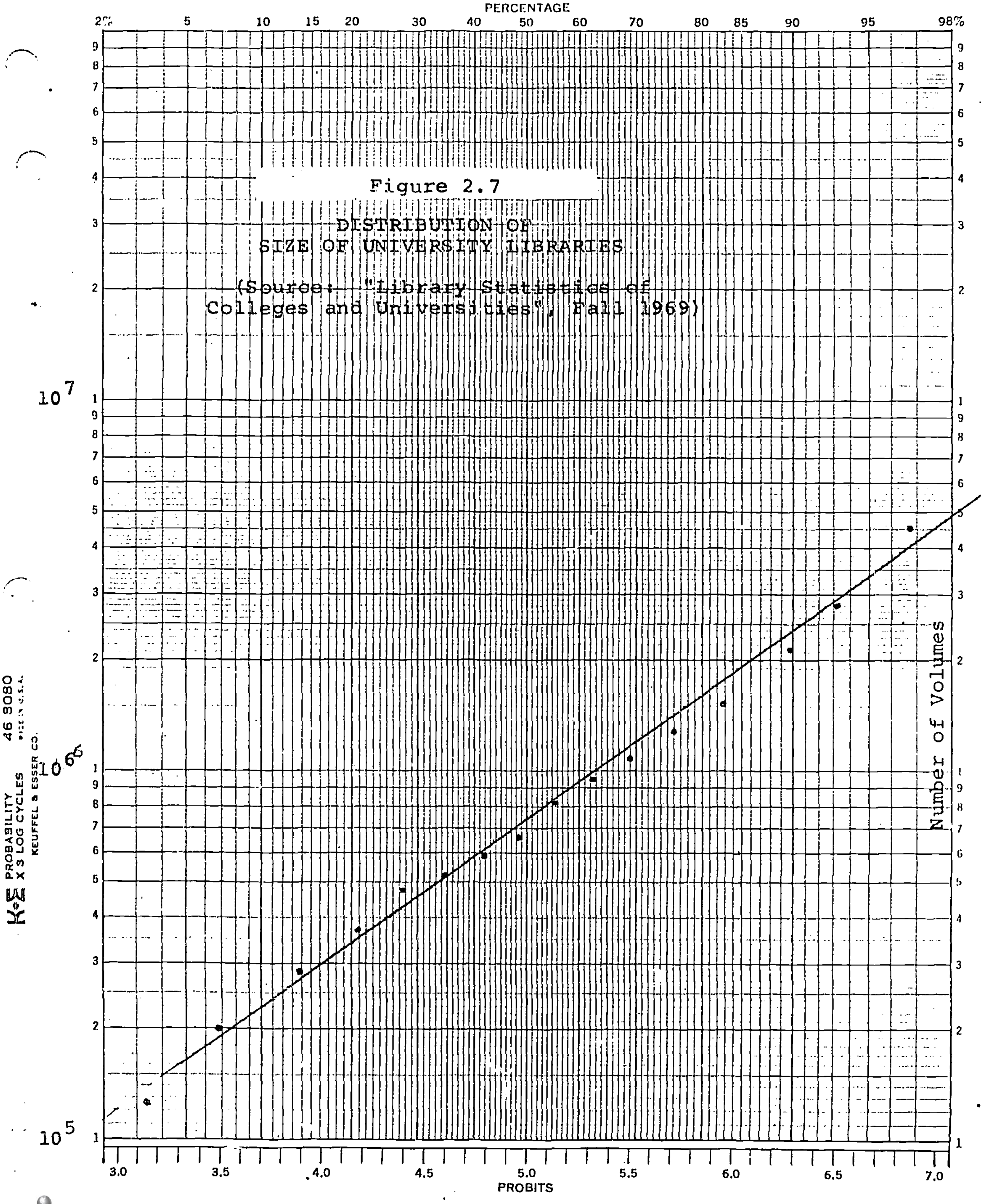
At this point we have established the mean size and distribution of size for book based bibliographic entities ranging in average size from about 30 characters up to 620 billion characters, entities which differ in size by a factor of 20 billion. Our immediate task is to demonstrate that there is a simple and reasonable model which encompasses the entire range of bibliographic entities in a systematic way, relating those of one size to those of another in a uniform and unvarying manner.

In order to proceed, recall that the book title, table of contents, index, and text are four bibliographic units of increasing average size; let us say that they belong to levels 1,2,3,4 respectively. Let Y_n stand for the base 10 logarithm of the average size of the units belonging to level n; Figure 2.8 displays the points whose coordinates are (n,Y_n) for n = 1,2,3,4, and also the point $(8,Y_8)$ where Y_8 is the base 10 logarithm of the mean size of a university library, and the point $(7,Y_7)$ where Y_7 is the base 10 logarithm of the mean size of a two-year college library, obtained by analyzing the first 206 two-year college libraries listed in Reference (3); this procedure is biassed, leading to a slightly high estimate of the mean size of two-year college libraries because the State of California dominates the initial part of the list both in number of two year colleges and in the size of their libraries,

Figure 2.7

DISTRIBUTION OF
SIZE OF UNIVERSITY LIBRARIES

(Source: "Library Statistics of
Colleges and Universities", Fall 1969)



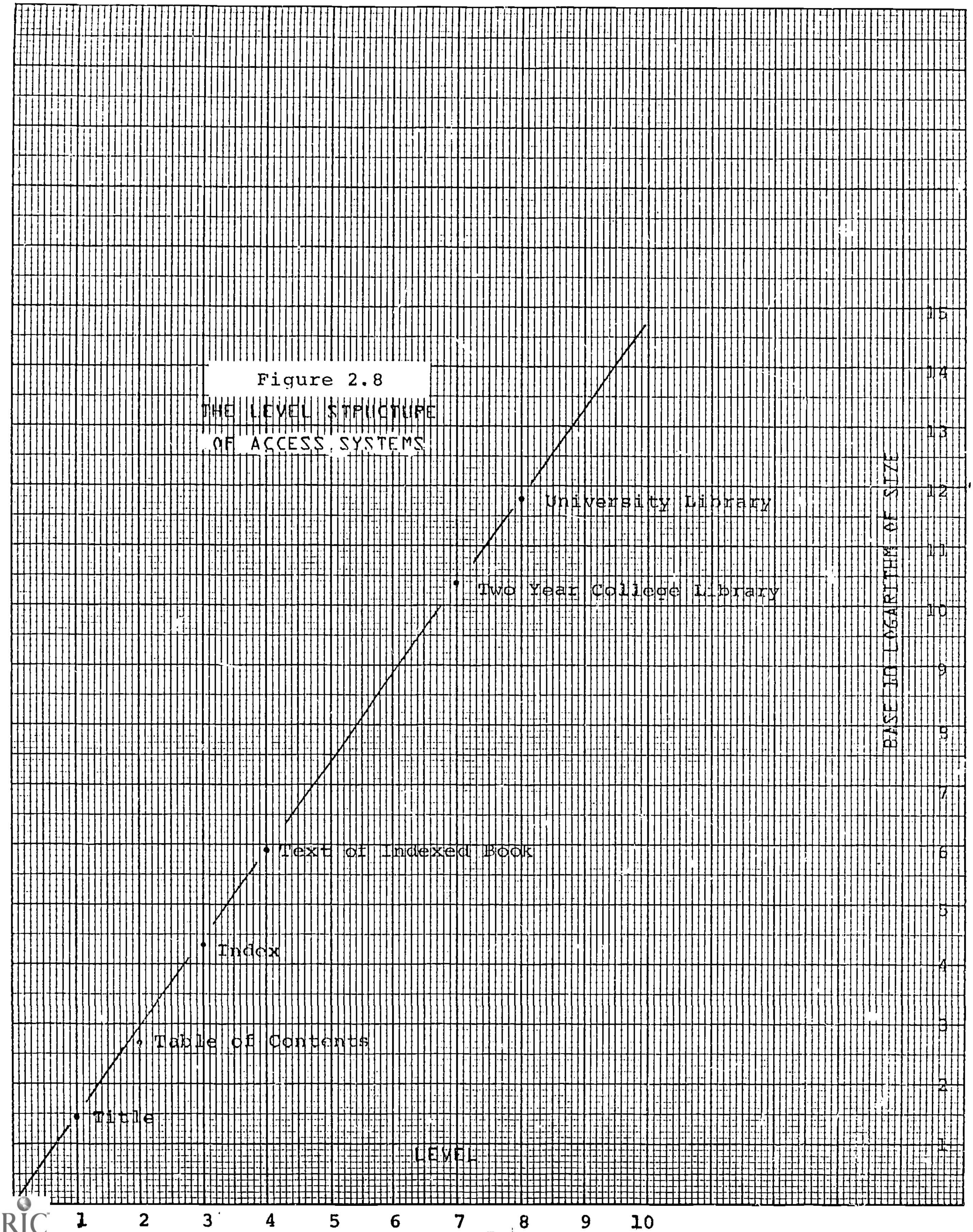


Figure 2.8
THE LEVEL STRUCTURE
OF ACCESS SYSTEMS

but analysis of the complete list in Reference (3), which is presently underway, will undoubtedly lower the mean size insignificantly from the value 29,912 volumes used to determine the corresponding point in Figure 2.8.

Figure 2.9 confirms that the size distribution of two year college libraries is lognormal and that the slope of the line representing the data on that graph is once again comparable with the slope of lognormal distributions presented in previous figures in this Chapter.

Inspection of Figure 2.8 may lead the reader to wonder whether levels 5 and 6 correspond to naturally occurring collections of books; we think that level 5 corresponds to general encyclopedias and level 6 to personal libraries, but we have not ventured to include calculations based on these hypotheses because of the difficulty of amassing reliable and comprehensive statistical information in their support.

The points in Figure 2.8 evidently lie very nearly on a straight line. This means that the mean size, s(n), of the bibliographic units comprising the n-th level is related to n by an equation of the form

$$s(n) = a10^{bn} \qquad (2.1)$$

where a and b are constants. It is natural to suppose that a = 1 so that level 0 corresponds to the single character; we will examine the data given in Figure 2.8 and Table 2.2 which corresponds to it to see if it is consistent with this desirable and simplifying hypothesis. By a standard application of the

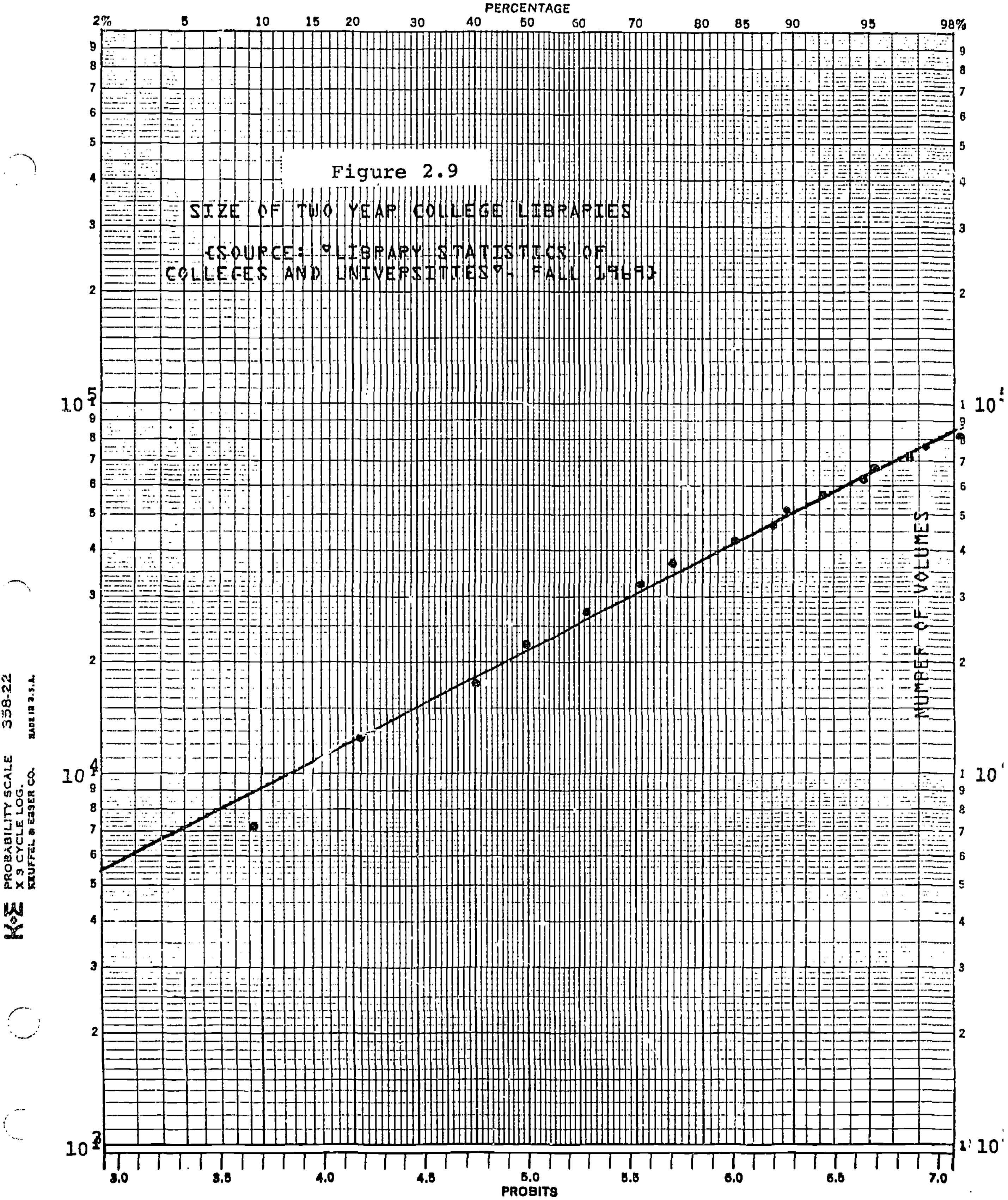


Figure 2.9

SIZE OF TWO YEAR COLLEGE LIBRARIES

{SOURCE: "LIBRARY STATISTICS OF COLLEGES AND UNIVERSITIES", FALL 1968}

Table 2.2

SIZE IN CHARACTERS

OF VARIOUS BIBLIOGRAPHIC UNITS

Unit	Level	Size	Log_{10} of Size
Title	1	28.15	1.44948
Table of Contents	2	505.	2.70329
Index	3	21293.	4.32710
Text of Book	4	819600.	5.91360
Two Year College Library	7	24528169200.	10.38966
University Library	8	620735534400.	11.79291

statistical F-test, as described for instance in Ref. (4), it is easily shown that the data does not contradict the hypothesis that a = 1 in eq. (2.1) at the 5% confidence level; this means that the least squares best fitting line for the points in Figure 2.8 does not differ significantly from that line which is constrained to pass through the origin of the coordinate system and also minimizes the sum of the squares of the deviations from the data points. This latter line corresponds to a relation of the form

$$s(n) = 10^{bn} \qquad (2.2)$$

relating the mean size of bibliographic units to their level. Carrying out the least squares minimization for a function of this form on the logarithms of the data leads to the line drawn in Figure 2.8 which corresponds to the equation

$$s(n) = 10^{1.47247n} = (29.68)^{n}. \qquad (2.3)$$

The constant 29.68 is an estimate of the fundamental constant determining the level structure of the bibliographic units considered above. More extensive data will no doubt result in the modification of this value, but it can be said with certainty that the fundamental constant is approximately 30, and perhaps may be identifiable with $(2e)^2 = 29.54...$, where $e = 2.718...$ is the mathematical constant denoting the base of the natural logarithm system.

This is our first main result:

> The average size of the bibliographic units *title*, *table of contents*, *index*, *monograph*, *two year college library*, and *university library* are powers of a fixed constant K whose value is nearly $(2e)^2$.

If it could be shown that the mean size of an encyclopedia is approximately K^5 and that of a personal (or perhaps a library reference sublibrary) is about K^6, then it could be asserted that the natural bibliographic units are *equispaced* when measured by the logarithm of their size; the current state of knowledge only permits us to assert that this is so for levels 1 through 4 and also for the separation of levels 7 and 8.

The previous argument suggests that the notion of *level* be introduced more generally. Therefore define the *level* of a given information base to be the integer closest to the logarithm of its size (the latter measured as usual in characters) to the base K; moreover, if a system of level K provides access to an information store of level n, then define the *order of access* provided by the access system as (n-k). Thus an index provides access of order 1 (=4-3) to the monograph it accompanies, and similarly the table of contents and title provide access of order 2 and 3 respectively to the book with which they are associated. We will later find that a library card catalog provides access of order 2 to the library archive but unfortunately it occupies a physical volume which could provide order 1 access to the collection.

Thus far we have principally concerned ourselves with the mean value of the various size distributions that have been examined, and have thereby shown that there is a simple and uniform relationship which connects the smallest of the natural units to the largest. We must now take up the question of the extent to which the mean characterizes the distributions that occur. The figures displaying the various distributions at the same time provide powerful evidence that all of the distributions are lognormal. The elementary form of the lognormal function, which is what occurs here, depends on two parameters--the lognormal mean and the lognormal standard deviation; if these parameters are known, then the usual mean value of the distribution can be determined and conversely, if the lognormal standard deviation and the usual mean are known, the lognormal mean and hence the lognormal function itself are completely determined (*cp.*Chapter III). From this it follows that if the lognormal standard deviation of the various distributions of interest are all essentially equal, then the associated lognormal functions are in reality determined by the mean value, that is, by the *level*, of the distribution. We shall show that this is indeed the case. Table 2.3 lists the lognormal standard deviation of the six distributions that have been described thus far.

Table 2.3

LOGNORMAL STANDARD DEVIATION

Unit	Level	Lognormal S.D.
Title	1	0.19
Table of Contents	2	0.30
Index	3	0.44
Monograph	4	0.23
Two Year College Library	7	0.29
University Library	8	0.36

There is evidently not much variation of the lognormal standard deviation as the level changes from a distribution whose typical size is about 30 characters to one whose typical size is about 600 billion characters and in particular what variation there is does not seem to have a trend. Based on the data contained in Table 2.3 we assert that the *lognormal standard deviation is essentially constant throughout the entire range of bibliographic interest*, and consequently *the distributions of size of the various bibliographic units are determined by the level of the unit*.

The lognormal standard deviation corresponds to the slope of the line defining the lognormal function for figures drawn on lognormal probability graph paper such as Figures 2.1-2.7 and 2.9 are. The underlined statement in the previous paragraph is the analytical version of the geometrical assertion that the lines representing all of the distributions are nearly parallel. We show to what extent this is so in Figure 2.10 which displays the distributions for all six levels; the variation of slope is indeed not great. The mean value of the standard deviations listed in Table 2.3 is 0.30, which may be conveniently adopted as an estimate of the level-independent lognormal standard deviation.

The assertion that the distribution of a variable x is lognormal is equivalent to stating that the distribution of log x is the normal (Gaussian) distribution. Here 'log' denotes the logarithm with respect to any conveniently chosen base. The graph of a normal distribution is the well known 'bell-shaped curve'. The level-structured lognormal distribution model of access systems described above can be equivalently viewed as a level-structured model for the logarithm of the size of bibliographic units such that the mean of the logarithms of the various levels are equally spaced and the associated distributions are normal, as shown in Figure 2.11 for levels 1-3. From that figure one also sees that the several bell curves have little overlap; this corresponds to the relative horizontality of the lines in the previous Figure 2.10 which is another way of stating that the lognormal standard deviation is a small number. The converse possibility, which fortunately does not occur, is that the lognormal standard deviation be relatively large with the consequence that the normal distributions like those illustrated in Figure 2.11 would possess a large degree of overlap with the overall appearance of gentle waves uniformly spread over a sea rather the sharply defined and separated peaks and valleys that Figure 2.11 so clearly exhibits. What this means is that the notion of level for bibliographic units makes sense; almost all units of some given type are of a size that is closer to the level of that type than to any other level. For instance, from Figure 2.10 we can read that fewer than 0.05% (sic!) of the Tables of Contents are so large as to lie (in logarithmic measure) closer to level 3 (Indexes) than to level 2 (Tables of Contents); similarly, fewer than 0.2% of the Two Year College Libraries are so large that they lie closer (in logarithmic measure) to the average size of a university library than to the average size of a two-year college library.

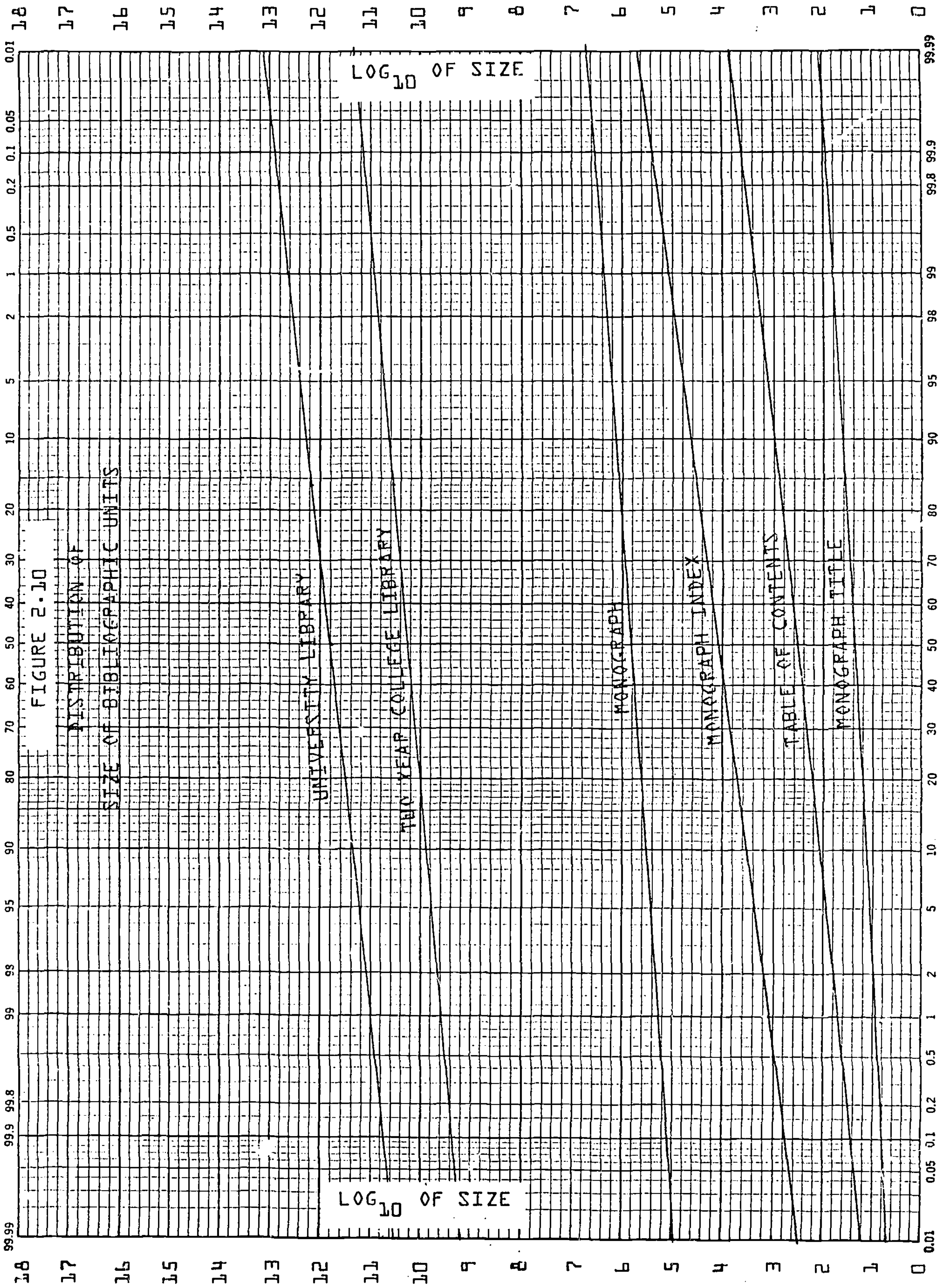


FIGURE 2.10
DISTRIBUTION OF
SIZE OF BIBLIOGRAPHIC UNITS

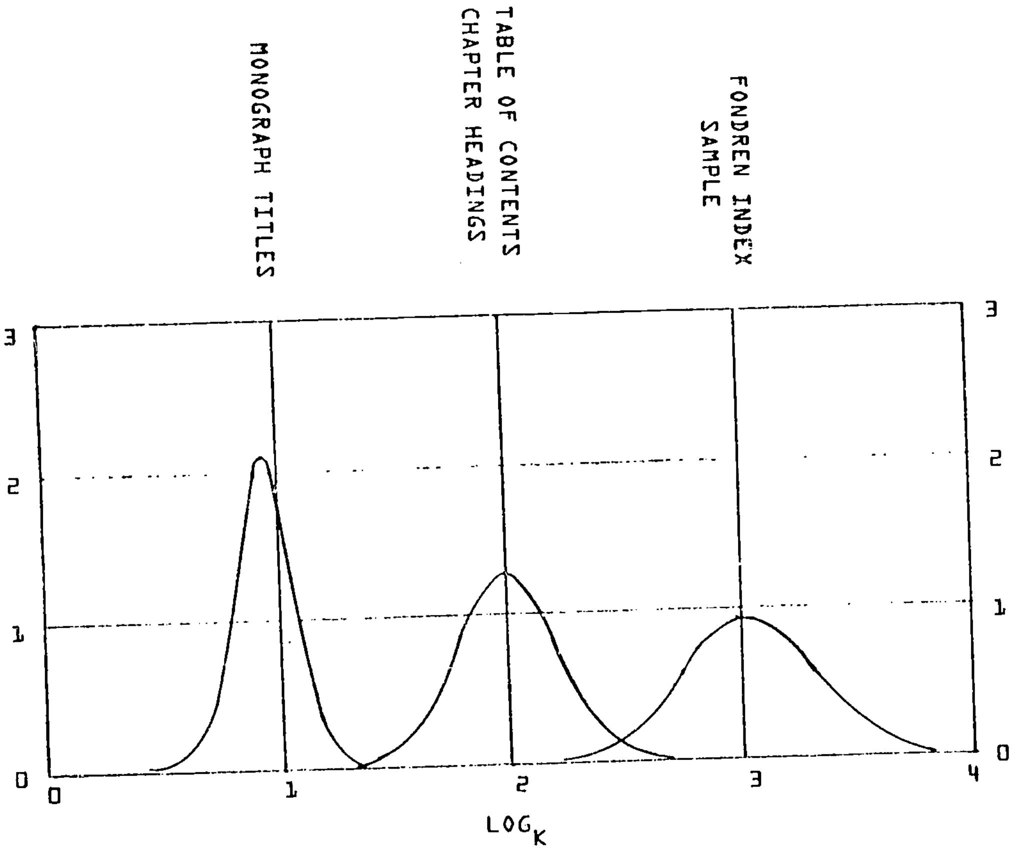


FIGURE 2.11

SOME ACCESS DISTRIBUTIONS IN LOGARITHMIC VARIABLES

These observations suggest that the notion of boundary separating two adjacent levels should be introduced as that size corresponding to half integer values of the level. More precisely, with level n and size s(n) related as in eq. (2.2), we say that the size s(n+1/2) is the boundary size between s(n) and s(n+1), and that (n+1/2) is the boundary between level n and level (n+1).

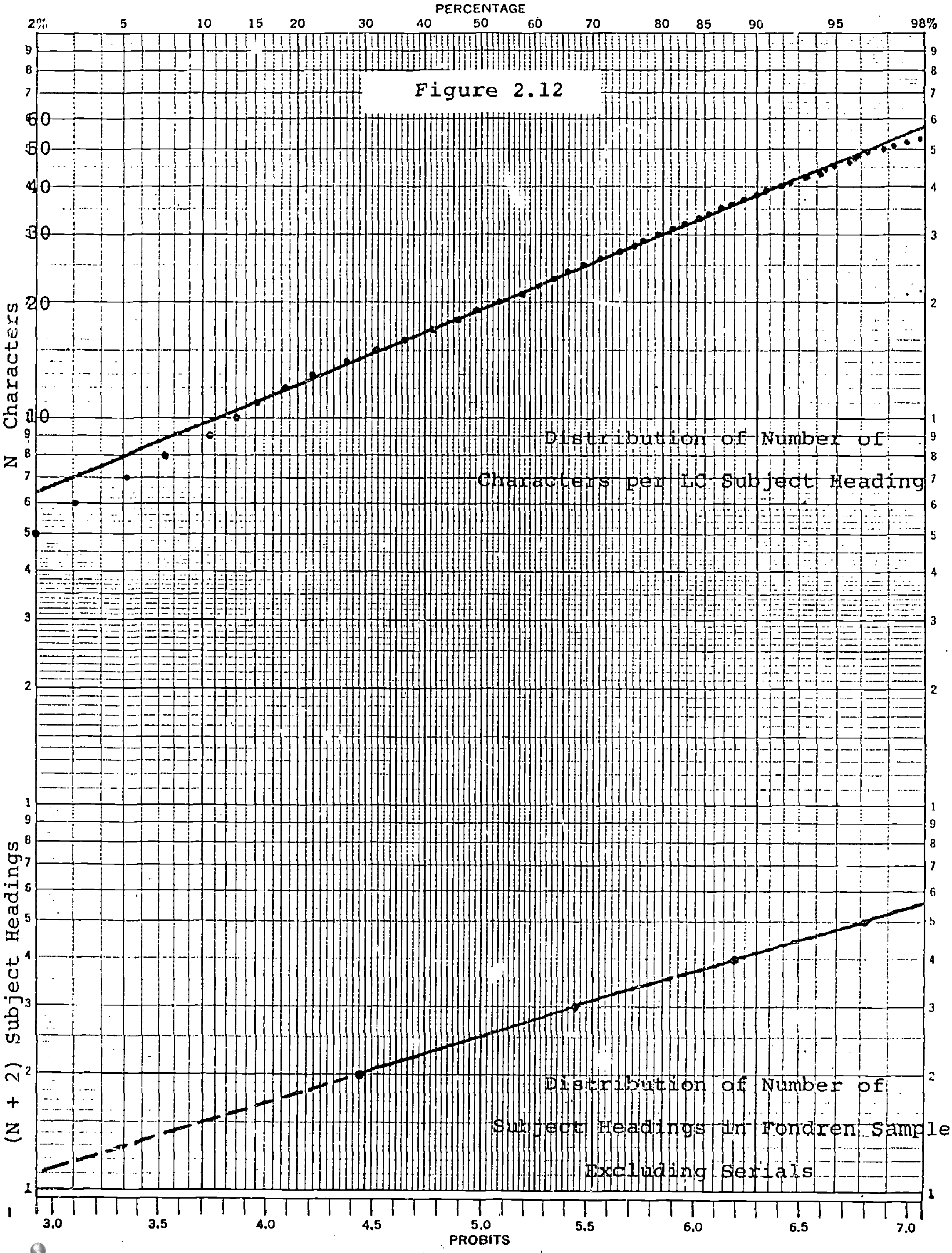
With this notion in hand it becomes possible to analyze a bibliographic item in order to determine if its size coincides reasonably with its 'proper' size, i.e., with the level of that type of bibliographic unit; from its size s compute $\log_K s$ and compare this number with the appropriate bibliographic unit level n to see whether $\log_K s$ lies within ± 1/2 of n; if it does not, then we may assert that the item of size s is either too large or too small. There will of course be specific exceptional instance for which the size of the unit is indeed 'proper' although not consistent with the statistically typical behavior for items of its bibliographic type, but the designer or evaluator of information access systems and/or information bearing data bases should, we think, warily approach the question of the size of a system from this point of view.

The access model presented in this chapter is not restricted to the book and its subsystems and supersystems. There is considerable evidence that it reflects universal properties of information stored in written English form, and, in a slightly generalized version, may be still more broadly applicable to the analysis and modeling of other types of information systems such as those associated with the modalities of sensory perception. These wide ranging and difficult issues cannot be examined here in a serious way; moreover, we do not yet have sufficient data upon which a definitive report can be based. Some of the intriguing vignettes that are most directly related to information presented in forms analogous to, if superficially distinct from, the book information system hierarchy explored above may nevertheless prove helpful for the reader.

First consider the size relationships of component units of the serial publication archive. We have studied the mathematics journal subarchive with the following results. For 7445 papers reviewed in volume 36 of Mathematical Reviews (published in 1968), the mean length of an abstracted paper is 13.8 'pages'; here 'page' refers to the myriad distinct page sizes and formats used by the 800-odd distinct journals reviewed by Mathematical Reviews. Bearing this in mind, and noting that we have not attempted to directly determine the mean number of characters per page of mathematics text nor the effect of the numerous special symbols which extend the normal type font, use of our previous estimate of 2400 characters per page of text yields the estimate of 33,120 characters per mathematics paper; hence such a paper is of level 3. The mean length of an abstract in Mathematical Reviews is easily estimated to be about 1081 characters. Therefore the size of the average mathematics paper is 30.6 times the size of the average abstract. Division of the estimated size of an abstract by K = 29.54 gives 36.59 characters, which is about the size of the average mathematics journal paper title and is of course quite close to the level 1 mean of 29.54 characters. We conclude that journal papers in mathematics are structured in a manner which is consistent with the general model proposed for books.

Next consider a more complex example which refers directly to the access problem. It is usual to find so-called "subject headings" at the foot of library catalog cards which are intended to provide cross reference access to subject areas other than those associated with the class number of the item corresponding to the catalog card. There are nearly 93,000 subject headings in the Library of Congress Subject Headings, seventh edition (1966). A uniform 1/66 sample drawn from an alphabetized list of these headings shows that the mean number of characters per subject heading is 22.3, which is not remarkably close to K = 29.54. However, the distribution of subject headings per catalog card as determined from an analysis of the Fondren Sample has a mean of 1.2 headings per card; if the distribution of subject headings per card is independent of the distribution of characters per subject heading, then the mean number of subject heading characters per catalog card, including the associated ordinals and interword space characters, will be the product of the means of the component distributions, which is 29.16. Hence the collection of subject headings per card provides about the same level of discrimination above the one-letter Library of Congress class in the mean that is provided by the title. Considering the distributions of characters per subject heading and subject headings per card leads to the lognormal functions shown in Figure 2.12; we conclude that the subject heading access mechanism is consistent with the level structured model and it belongs to level 1.

Figure 2.12



Distribution of Number of Characters per LC Subject Heading

Distribution of Number of Subject Headings in Fondren Sample Excluding Serials

The phenomenon that the mean value of the size of adjacent access levels are in the ratio of about 30 to 1 is not confined to access systems associated with written natural language archives. Consider ALTEXT, a contemporary text-processing higher level (macro expander) computer language [5]. Such a language consists of computer instructions which have two parts: a *generic instruction* such as the GOTO of FORTRAN which *specifies the general* function of the instruction, and certain other more particular components which contain the details of data location and transfers of control. The *implementation* of a higher level computer language instruction consists of a sequence of one or more "machine language" or "assembly language" instructions; the advantage of the higher level language is that it frees the programmer from the burden of keeping track of numerous housekeeping details concerning the location and manipulation of the data at the cost of lower (local) efficiencies of execution. This is another way of stating that the higher level language instructions act as an access system for the sequences of assembly language instructions that are their implementation.

With this preamble in mind, one can examine the number of assembly language instructions required to implement each of the distinct generic higher level language instructions. For the generic instructions of ALTEXT, the mean number of assembly language instructions per ALTEXT "macro" is 30.32 (including implementation of the "ALTEXT macro" which provides the interface with the operating system of the implementing computer) for implementation on the IBM 360/30 computer. Figure 2.13 confirms in a rather startling way that the distribution of implementation size is lognormal; hence we conjecture that the level structured access model will probably find significant application in the design of computer languages.

That the structure of many types of linguistic units is lognormal has long been known and abundantly verified. The lognormality of word length statistics was discovered at least as early as 1887 by Mendenhall [6] and was subsequently studied, along with sentence length distributions, *inter alia*, by Yule [7], Williams [8], and Herdan [9]. *Yule* computed the sentence length distributions for a number of samples of written English and although he did not notice their lognormality himself, Williams did test this hypothesis on Yule's data and on more he gathered himself. More extensive data has been collected by Kucera and Francis [10] but care must be exercized to insure that it is partitioned into homogeneous subject and/or author classes before attempting to study the lognormality of the statistics; the problem of describing the structure of inhomogeneous data, which amounts to studying how distinct lognormal distributions combine, is relatively complex. Moreover, much of the Kucera and Francis data refers to printed materials that are unlikely to form an active part of an archival library collection; it is heavily weighted with fiction and press coverage.

Herdan |9| analyzed 80,000 words of telephone conversations collected by French, Carter and Koenig of the Bell Telephone Laboratories and concluded that (phonetic) word length is lognormally distributed. An indication that the parameters of these linguistic distributions are relatively insensitive to variations in language vocabulary and to whether the written or spoken form is used is provided by Figure 2.14 which shows nearly parallel lines representing the Herdan telephone conversations and Mendenhall's analysis of 1000 words from Shakespeare's works (as represented by Williams).

These examples and others too numerous to report here prompt us to speculate that the occurrence of the lognormal distribution is fundamental to all human information processing activities. In this regard we distinguish two types of activities: those that process direct sensory impressions that are received through the sensory organs, and those that process coded information such as is represented by linguistic codes. In the latter instance the directly perceived data arrives via the sensory organs but the essential content is unrelated to the particular code used for its transmission. Although there may be important differences between the internal mechanisms that process these two types of information, there are at least two characteristics that the two types of input information share: the quantity of information that passes through the processing system is very large and the system must be capable of responding to inputs whose size vary greatly. The first condition requires that the information processing system be able to compress (with information loss) the vast amount of data passing through it so as to be enabled to retain for future use a much smaller but characteristic subset of it; in other words, the processing system must function as an access system to the information passing through it. The second condition suggests that some functional transformation must be applied to the input sensory information in order to reduce its extended range to a smaller one more conveniently handled by the neural network; for example, there has long been evidence (which is reflected by the 'decibel' scale of measurement) that the subjective response to the stimulus provided the ear by acoustic energy varies as the logarithm of the input energy.

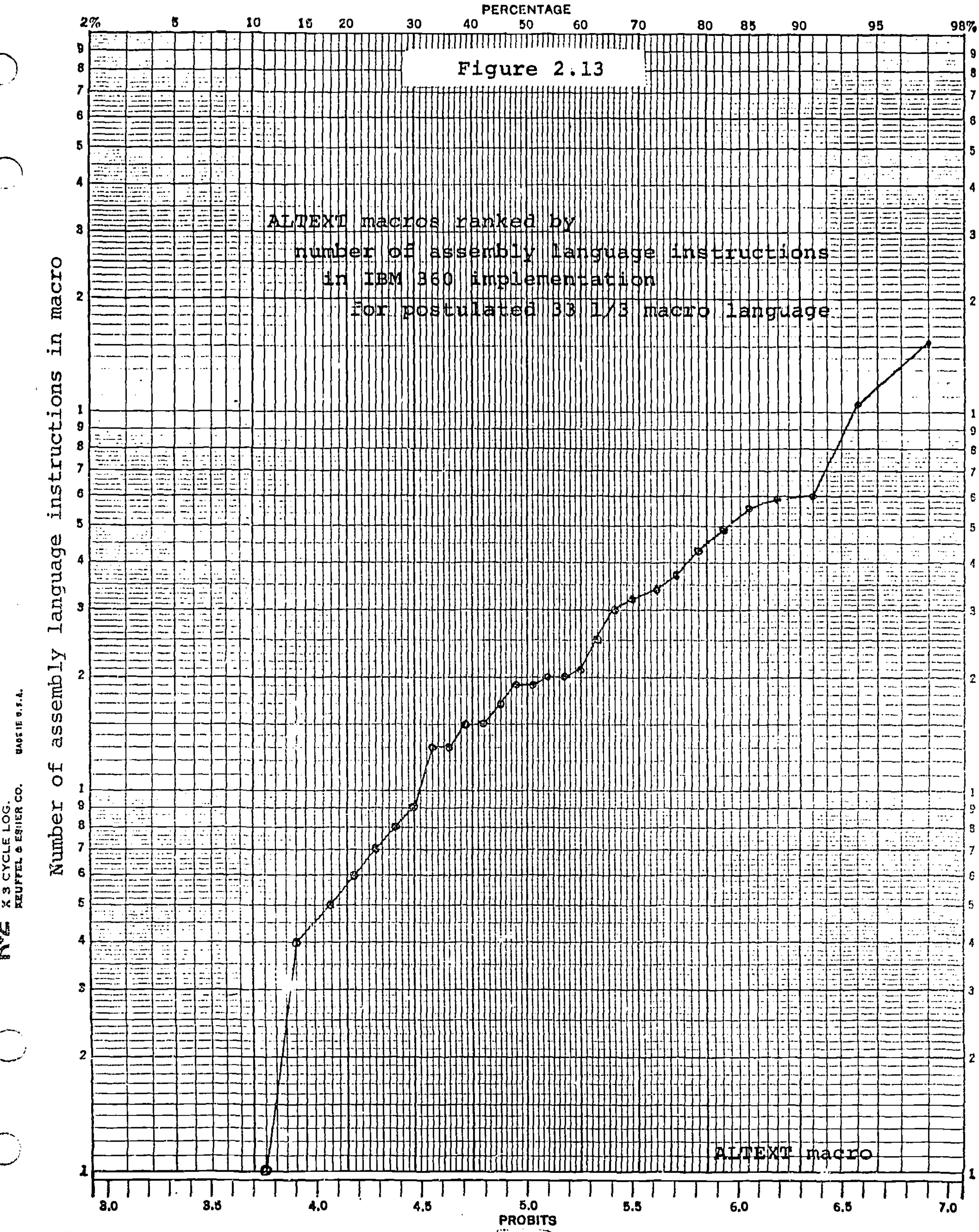
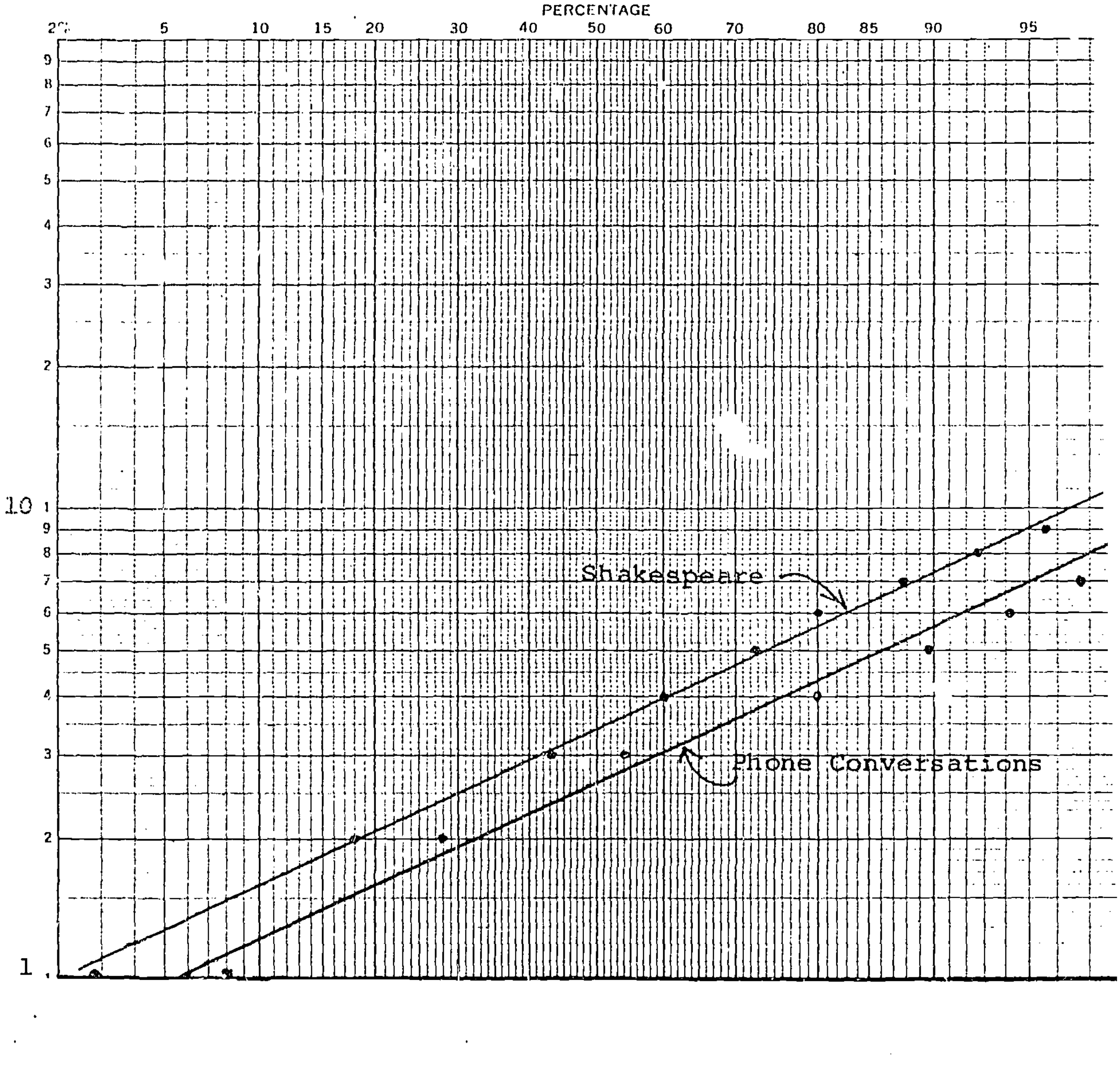


Figure 2.13

ALTEXT macros ranked by number of assembly language instructions in IBM 360 implenentation for postulated 33 1/3 macro language

Figure 2.14

Word Length Distributions
(in characters)



Generally, there are three reasons for making a scale transformation in analyzing data (e.g., see Tukey [11]):

1. To linearize the relation between two variables.

2. To normalize the underlying probability distribution.

3. To stabilize the variance.

Although in most applications any one of these results would provide sufficient reason for introducing a particular transformation, it is not uncommon to encounter situations where the transformation is originally introduced for one reason and subsequent analysis shows one or both of the remaining desiderata have also been achieved.

In this context it is illuminating to study the work of the nineteenth century experimental psychologist G. Fechner [12]. He made the important observation that the ability of the human to respond a stimulus is proportional to the mean level of the stimulus. That is, if an individual can just sense a difference of, say, one unit when the mean level of stimulation is 10 units, then he will also just be able to detect a difference of 2 units when the mean level is 20 units. This multiplicative property of the just noticeable difference led him to introduce the logarithm function in order to stabilize the variance, i.e., make it constant throughout the range of perception. He then conjectured that the function relating subjective response to the transformed variable--the logarithm of the stimulus--is a linear function, thus arriving at the celebrated (and once again hotly debated) 'Law' of Weber and Fechner. The reader will observe that the logarithm of the size of bibliographic units stabilizes the variance of the distributions of these units throughout the entire range of 'bibliographic perception'. This certainly makes it tempting to inquire whether the Weber-Fechner 'Law" might not be merely an approximation to some more accurate description of the underlying functional transformation governing sensory perception. This question has received considerable attention in recent years and notable contributions have been made, principally by Stevens (e.g., [13]), who has generalized the logarithmic Weber-Fechner transformation so that response is some power of stimulus; that this change actually constitutes a generalization becomes clear when it is noted that the integral of 1/x is log x whereas the integral of any other power of x is again a power of x; in this sense the logarithm is the limit of power functions (see Dolby [14]). The relationship between linguistic and hence bibliographic units and these psychophysical questions has been remarked by several workers, most notably perhaps by Fairthorne [15]; Zipf's 'Law' [16] in its integrated form is just the Weber-Fechner logarithmic relation, and Mandelbrot's [17] generalization of Zipf's function corresponds--indeed, it is identical to--Steven's power function. These questions will be taken up from a more mathematical standpoint in the next chapter with the intent of showing how they can be derived, following an argument essentially due to Mandelbrot, from elementary considerations from information theory, and, of more importance for our purposes, that a slight extension of this argument generalizes the Weber-Fechner-Zipf-Stevens-Mandelbrot functions to the lognormal distribution. For as the extensive bibliographic data assembled in the earlier parts of this chapter show, it is the lognormal function that in fact describes reality.

References

1. Dolby, J. L., V. Forsyth, and H. L. Resnikoff, Computerized Library Catalogs: Their Growth, Cost and Utility, M.I.T. Press, Cambridge, 1969.

2. Dolby, J. L. and W. J. Jones, "The Measurement of Composition Practice", in Advances in Computer Typesettting, Institute of Printing, London, 1966.

3. Price, Bronson, Library Statistics of Colleges and Universities, Fall 1969, Data for Individual Institutions, U. S. Offfice of Education, Washington, D. C., 1970.

4. Youden, W. J., Statistical Methods for Chemists, John Wiley & Sons, New York, 1951.

5. Dolby, J. L., W. E. Houchin, Roger Stark, and H. L. Resnikoff, Non-Numeric Programming Language Studies: ALTEXT II, Final Report to U.S.A.F. Office of Scientific Research, R & D Consultants Co., Los Altos, California, 1970.

6. Mendenhall, T. C.,"The Characteristic Curves of Composition", Science, 9,(214,supplement) (1887), 237-49

7. Yule, G. U., "On Sentence-Length as a Statistical Characteristic of Style in Prose", Biometrika 30(1939), 363-84

8. Williams, C. B., "A Note on the Statistical Analysis of Sentence-Length as a Criterion of Literary Style", *Biometrika*, 31(1940), 356-61.

9. Herdan, G., "The Relation between the Dictionary Distribution and the Occurrence Distribution of Word Length and its Importance for the Study of Quantitative Linguistics", *Biometrika*, 45 (1958), 222-8.

10. Kucera, Henry and W. N. Francis, *Computational Analysis of Present-Day American English*, Brown University Press, Providence, 1967.

11. Tukey, J. W., "On the Comparative Anatomy of Transformations", *Annals of Mathematical Statistics*, 28(1957), 602-32.

12. Fechner, G. T., *Elemente der Psychophysik*, 1860.

13. Stevens, S. S., "Neural Events and the Psychophysical Law", Science, 170(1970), 1043-50.

14. Dolby, J. L., "A Quick Method for Choosing a Transformation", *Technometrics*, 5(1963), 317-25.

15. Fairthorne, R. A., "Empirical Hyperbolic Distributions (Bradford-Zipf-Mandelbrot) for Bibliometric Description and Prediction", *Journal of Documentation*, 25(1969), 319-43.

16. Zipf, G. K., *Psycho-Biology of Language*, Houghton Mifflin, 1935.

17. Mandelbrot, B., "An Information Theory of the Statistical Structure of Language", *Proceedings of the Symposium on Applications of Communication Theory, London, September 1952*, Butterworth, 1953, 486-500.

CHAPTER III

MATHEMATICS OF INFORMATION DISTRIBUTIONS

MATHEMATICS OF INFORMATION DISTRIBUTIONS

This chapter is devoted to the mathematical study of some of the distributions that arise naturally in the study of information systems. It will necessarily be more demanding of the reader's mathematical knowledge than the remainder of the book and has therefore been written so as to permit the reader to pass immediately to Chapter IV without loss of continuity. We believe, however, that the significance and implications of the level structured model of access systems presented in Chapter II cannot be fully understood unless the relationship of that model to other competing models, extant and potential, is made clear. Moreover, the most powerful theoretical arguments for the appearance of the lognormal distribution in the model structure come from information theory and its mathematical apparatus, so there is really no way to avoid these technical considerations.

We will be principally concerned with two distributions--the Zipf and the lognormal. The former is also frequently associated with the names Estoup (1), Bradford (2), and more recently Mandelbrot (3,4). Zipf rediscovered and popularized the observation that the ranked frequency distribution of words in natural text corpora is essentially of the form

$$y = cx^{-s} \qquad (3.1)$$

where x denotes the rank and y(x) the frequency of occurrence of the word of rank x; here c and s are constants selected to fit the data as nearly as possible and which therefore are characteristic of the text corpus and to some extent the language from which it is drawn; Zipf only considered the case s = 1. Figure 3.1, taken from Zipf (5), exhibits such distributions.

Distributions of the type (3.1) occur in other fields, associated, for special values of s, with Pareto (6) in economics, Lotka (7) in what might be termed 'sociological mathematics', and more recently De Solla Price (8), and no doubt in numerous other contexts as well.

That the Zipf "law" is taken seriously, not just considered as an accidental quirk of the data to be remarked upon and ignored, is attested by the variety of publicatons that dispute, modify, and reduce it to a triviality. Mandelbrot showed that a slight generalization of the

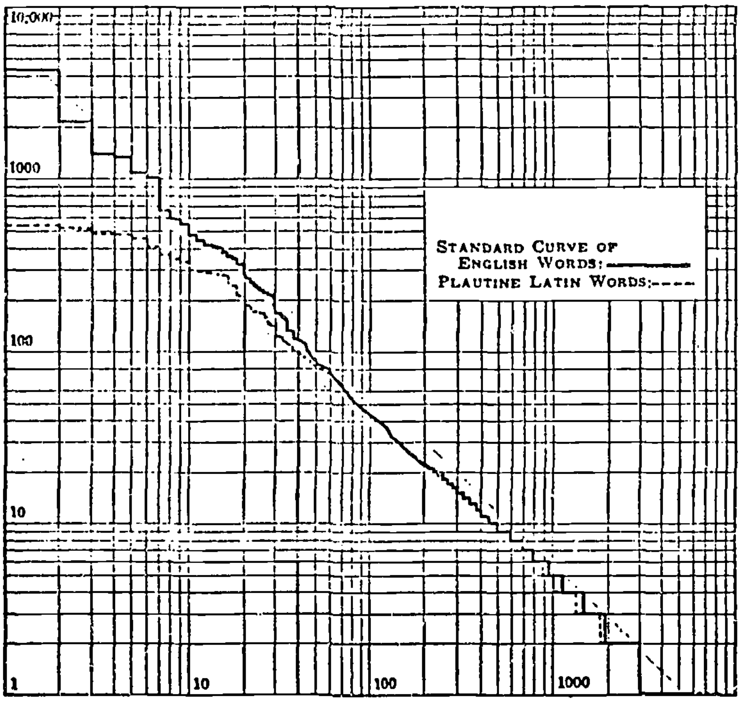

10,000
1000
100
10
1
10
100
1000
STANDARD CURVE OF ENGLISH WORDS: ———
PLAUTINE LATIN WORDS:----

Figure 3.1

Zipf law, in better agreement with the data, is a consequence of elementary arguments and reasonable hypotheses about the effort required for the efficient transmission of information; in this sense he is a bulwark for both the proselytizers and trivializers of Zipf since his arguments convincingly show that the nature of the distribution has nothing to do with special properties of language that distinguish it from a variety of other processes that extremize some function representing the degree of organization of the process in a statistical sense. Mandelbrot was quick to point out the connection, which is more than merely formal, between his result and the mathematical methods used to derive it, and the derivation of the partition function in statistical mechanics; cp. Schrodinger (9).

Despite the considerable research efforts that have gone into understanding and improving the relation of Zipf, there are significant discrepancies between the Zipf-Mandelbrot predictions and the observed data for large samples of words drawn from natural language text corpora and for other data collections as well. There are theoretical difficulties too: Yule (10) observed that the sum of the frequencies predicted by the Zipf-Mandelbrot distribution (with s = 1) is not finite, which implies that there must be a significant deviation from this distribution for large values of the variable. This kind of difficulty is not as easily brushed aside as disagreement with the data can be, for it entails an unknown mechanism which determines that range of the variable for which the distribution must be modified as well as the unknown modification itself, and leaves the researcher bereft of the argument that improved "experimental measurements" will modify the situation in any agreeable way. It is much easier to reconcile ill fitted observations, and their consequences are normally much more local in nature.

Nevertheless it has been found desirable to modify Zipf's Law in many ways to better fit the data. Mandelbrot's modification, based on his theoretical considerations, is

$$y = c(x - a)^{-s} \qquad (3.2)$$

where a is a small constant. Belonogov (11) found that the distribution

$$y = e^{-c(x-1)^k} - e^{-cx^k} \qquad (3.3)$$

describes the rank-frequency structure of printed commercial Russian. Good (12) is led to

$$y = c(x-a)^{-s(1 + by^{-1})}, \qquad (3.4)$$

with b a small constant; this form has (3.2) as a first approximation (because b is small) and also is responsive to Yule's criticism since the sum of the frequencies is finite. It is derived by including in the effort function (see below) a factor coresponding to the effort required to incorporate words of large rank in the inventory, and represents, in a certain sense, part of the system 'overhead'. Unfortunately, (3.4) is a complicated expression and Good's choice of overhead factor is in no way uniquely determined.

Other authors have turned to functions that are apparently quite different in order to more faithfully describe their data. Houston and Wall (13) described the distribution of term usage in manipulative indexes using the lognormal distribution, eliciting from Fairthorne (14) the remark that, in his view, Wall (15) (and presumably also Houston and Wall (13)) selected the lognormal only because the data was well fit by that distribution in the sense that the results plotted as a straight line on lognormal probability graph paper, but that they would also have done so on ordinary logarithmic graph paper because "segments of the tail of a Gaussian distribution are not readily distinguished from segments of a hyperbolic distribution"; by the latter he means the Zipf distribution. Certainly this remark applies in principle to the figures plotted on lognormal graph paper in the previous Chapter, but we will see that it is significant only when the variance of the lognormal distribution in question is large.

Carroll (16) has discussed the statistical problems associated with representation of the Standard Sample of Present-Day Edited American English (17) by lognormal distributions; there is in his work no hint of the formerly used Zipf approximation.

We think that there are two conclusions that should be drawn from this necessarily brief survey: first, the Zipf-Mandelbrot distribution does not adequately fit much data although it is well grounded in theory, and second, it is often difficult to distinguish lognormal approximations of data from Zipf approximations. They suggest that an intimate relation may exist connecting the Zipf and lognormal distributions, and, if this be true, that a 'derivation' of the lognormal from elementary principles along the lines of Mandelbrot's arguments may be possible.

In order to show that these hopes are indeed justified, we will present a derivation of the Zipf-Mandelbrot Law following the usual argument, and in effect following Schrödinger (9), although the notation and terminology there is of course quite different.

Consider information 'states' $S_1, S_2, \ldots, S_x, \ldots$ constituting some inventory, such as the words of language as they occur in some large text corpus, or a large random collection of monograph titles or indexes, ordered in some convenient fashion. Let $\varepsilon(x)$ denote the 'effort' ('energy') required to utilize state S_x in an access system ('communication system'), and denote by $p(x)$ the probability of utilization of S_x in the inventory. Following Shannon (18), the expected amount of information per unit expected effort is proportional to

$$I = -\sum p(x) \log p(x) \,/ \sum p(x)\varepsilon(x). \qquad (3.5)$$

If the access system is such that the expected amount of information per unit effort is maximized, then the probabilities $p(x)$ cannot be unrelated to the effort function $\varepsilon(x)$; maximization of I subject to the necessary restraint

$$\sum p(x) = 1 \qquad (3.6)$$

will determine the form of $p(x)$ for given $\varepsilon(x)$. Now maximization of (3.5) subject to (3.6) is equivalent to maximization of

$$-\sum p(x) \log p(x) \qquad (3.7)$$

subject to (3.6) and the additional restraint that the total effort

$$\sum p(x)\varepsilon(x)$$

is constant as well. The mathematical method of Lagrange multipliers provides the solution to this extremal problem in the following way: since the total probability (3.6) and the total effort are constant, the function (3.7) and the function

$$H = -\sum p(x)\log p(x) \quad + \quad (1+a_0)\sum p(x)$$
$$+ a_1 \sum p(x)\varepsilon(x) \qquad (3.8)$$

attain their maximum for the same functions p(x) of e(x), where a_0 and a_1 are arbitrary constants and the form $(1+a_0)$ has been chosen for later notational convenience. Now subject each p(x) to small differentiable independent functional variations, all the while keeping x fixed; H will assume its maximum where the derivatives $\partial H/\partial p(x)$ all vanish. This yields the simultaneous conditions

$$0 = \partial H/\partial\, p(x) = -(1 + \log p(x)) + (1+a_0) + a_1\varepsilon(x) \qquad (3.9)$$

which implies

$$\log p(x) = a_0 + a_1\varepsilon(x) \; . \qquad (3.10)$$

This is the fundamental relation connecting the effort function, which is presumed to be known, with the probability of occurrence of the state S_x. It remains to specify the effort function. Mandelbrot argued that, if the states S_x are words drawn from natural text and arranged in decreasing frequency of occurrence, then $\varepsilon(x)$ is proportional to log(x-a) with a some small constant. This hypothesis immediately leads to (3.2) with $c = e^{a_0}$ and $s = -a_1$. In order that the distribution decrease with increasing x, a_1 must be negative. (If $a_1 = 0$, the distribution degenerates into the uniform distribution, which can only apply to a finite range of the variable x.)

The idea underlying Mandelbrot's choice of effort function is perhaps most simply illustrated by recalling the ordinary use of positional notation to represent positive integers. If b is an integer greater than 1 and n is any positive integer, then n has a unique representation of the form

$$n = a_N b^N + \ldots + a_k b^k + \ldots + a_1 b + a_0$$

with a_k integers less than b but not negative. For instance, if b = 10 and n = 234, then

$$234 = 2 \cdot 10^2 + 3 \cdot 10 + 4.$$

By means of such an expression n can be identified with the sequence of numbers $a_N a_{N-1} \ldots a_0$. If b = 10 this correspondence is the usual decimal expression for n, while if b = 2, it is the binary expression. Such an expression for n requires N symbols each of which is selected from an inventory of b symbols (0,1,2,...,b-1). Evidently

$$N + 1 \geq \log_b n \geq N,$$

so the number of places required to express n in base b is approximately $\log_b n$, and approximately b $\log_b n$ selections suffice to specify an integer lying between 0 and n.

By coding the information specified by the states S_x as integers, this argument can be made to apply to the S_x themselves, leading to the Zipf-Mandelbrot distribution for words if that is what the states represent.

It must be recognized that more is involved in the distribution of information states than the simple matter of minimal coding; Good's argument mentioned above attempts to account to some extent for the effort required to add a state to the inventory, i.e., to learn a rare word. Therefore the effort function may not have the form proposed by Mandelbrot except in the simplest of cases, and it becomes necessary to investigate the probable nature of substitutes for it. It might be argued that in general a multiple of log (x-a) will constitute a good first approximation to $\varepsilon(x)$. This, and equation (3.10), suggest the introduction of the variables

$$u = \log (x-a) \qquad (3.11)$$

and

$$f(u) = \log p(x) . \qquad (3.12)$$

so that (3.10) becomes

$$f(u) = a_0 + a_1 \varepsilon(\varepsilon^u + a) = \varepsilon^*(u), \qquad (3.13)$$

defining the function $\varepsilon^*(u)$ which is more convenient to work with.

Mandelbrot's assumption for the effort function is, in this notation, simply that

$$\varepsilon^*(u) = a_0 + a_1 u \ . \qquad (3.14)$$

We will assume that $\varepsilon^*(u)$ can be expanded in a Taylor series about the point $u = 0$, that is,

$$\varepsilon^*(u) = \sum_{k=0}^{\infty} a_k u^k \ . \qquad (3.15)$$

For small values of u, $\varepsilon^*(u)$ will be well approximated by the first two terms of (3.15) if a_1 is not zero, leading to the Zipf-Mandelbrot Law; if a better approximation is desired, more terms must be taken from the series expansion of $\varepsilon^*(u)$. Suppose for instance that an approximation accurate through terms quadratic in u is used:

$$\varepsilon^*(u) = a_0 + a_1 u + a_2 u^2 \ .$$

Using (3.11) through (3.13), we obtain

$$\log \ p(x) = a_0 + a_1 \log(x-a) + a_2 (\log(x-a))^2;$$

the right hand side can be written as

$$-\log \ (x-a) + a_2 \left| \left(\log(x-a)+(1+a_1)/2a_2\right)^2 + \left(a_0/a_2-\left(1+a_1\right)/2a_2 \ ^2\right) \right|$$

so

$$p(x) = \frac{c \ e^{-\frac{1}{2}\left|\frac{\log(x-a)+(1+a_1)/2a_2}{\sqrt{-1/2a_2}}\right|}}{(x-a)} \qquad (3.16$$

$$= \frac{ce^{-\frac{1}{2}\left[\frac{\log(x-a)-m}{s}\right]^2}}{(x-a)}$$

with

$$c = e^{(4a_0a_2-(1+a_1)^2)/4a_2} \qquad (3.17)$$

which is the lognormal distribution with lognormal mean

$$m = (1 + a_1)/2a_2 \qquad (3.18)$$

and lognormal standard deviation

$$s = \sqrt{-1/2a_2} \qquad (3.19)$$

In other words, the parameters a_1 and a_2 which appear in the effort function $\varepsilon^*(u)$ are related to the parameters of the lognormal distribution defined by that effort function as follows:

$$a_1 = -(1 + \frac{m}{s^2}),\ a_2 = -1/2s^2\ , \qquad (3.20)$$

showing in particular that a_2 must be negative in order that the distribution correspond to a realizable system. Using these values for the constants appearing in the effort function yields

$$\varepsilon^*(u) = a_0 - (1 + m/s^2)\ u - u^2/2s^2\ , \qquad (3.21)$$

which shows that Mandelbrot's hypothesis is warranted when s and m are large in such a way that the quotient m/s^2 remains finite. If s is not large, then necessarily the simple logarithmic effort hypothesis is inadequate.

This observation reconciles Fairthorne's remark, quoted above, but it has the further reaching consequence that the question of whether the Zipf or the lognormal provides a 'better' fit to given data is really meaningless from this point of view; the Zipf is a special case of the lognormal and can therefore never provide a better fit by the implied metric than the latter. Moreover, the larger s is, the easier it will be to confuse the two distributions, since the general form will more nearly approach its specialization as s increases.

The derivation of the lognormal distribution given above is based on the measure, I, of information per unit effort defined by eq(3.5) which occurs in Mandelbrot's work and is also used by Good. It is, however, not the only reasonable measure of average effort. In fact, as W. E. Houchin has observed in a personal communication, the choice of

$$I^* = -\sum p(x) \left\{ \log p(x)/\varepsilon(x) \right\} ,$$

the expected value of the information per unit effort, in place of I (which is the expected value of the information per expected value of the effort) leads to the lognormal function in a more direct manner, free of the burdensome hypothesis that the effort function is quadratic in the logarithm of size or rank. For, arguing as before with I* in place of I, leads to the maximization of

$$H^* = -\sum p(x) \left\{ \log p(x)/\varepsilon(x) \right\} + a_1^* \sum p(x) + a_2^* \sum p(x)\varepsilon(x) ,$$

and therefore to the equations

$$0 = \varepsilon(x)\partial H^*/\partial p(x) = -(1 + \log p(x)) + a_1^*\varepsilon(x) + + a_1^*\varepsilon(x)^2 ,$$

with solution

$$\log p(x) = -1 + a_1^*\varepsilon(x) + a_2^*\varepsilon^2(x) .$$

If the effort is a logarithmic function of x, $\varepsilon(x) = \log(x-a)$, then p(x) is the lognormal function with lognormal mean

$$m^* = (1+a_1^*)/2a_2^*$$

and lognormal standard deviation

$$s^* = \sqrt{-1/2a_2^*} ;$$

these equations should be compared with eqs (3.18) and (3.19). If $m^* = 0$, then this lognormal function reduces to Zipf's original 'law' with exponent -1; if m^* and s^* are both large such that m^*/s^* is finite the general power function of Mandelbrot and Stevens results.

Recalling that $f(u) = \log p(x) = \varepsilon^*(u)$, (3.21) shows that the graph of $f(u)$ vs. u, that is, of $\log p(x)$ vs. $\log(x-a)$, will be a parabola if $p(x)$ is lognormally distributed, whereas it will be a straight line if $p(x)$ is distributed according to Zipf's Law. It is instructive to examine some examples.

Figure 3.2 displays the graph of the frequency of the most frequent ordered pairs of words drawn from English language text as a function of their rank, drawn on logarithmic graph paper. The data was derived from the corpus which constitutes the Standard Sample of Present-Day Edited American English (17). It approximates a straight line without any considerable evidence of global curvature thereby supporting the hypothesis that the effort function $\varepsilon^*(u)$ is linear and the consequent distribution Zipf. The arguments that are usually applied to justify the Zipf approximation for the distribution of frequency ranked single words would seem to apply equally well to this case, and therefore the arguments of Carroll (16) concerning the problems associated with the extraction of finite samples from theoretical distributions of lognormal type are also probably valid here, which helps to explain the bending of the curve in the direction of low frequencies for large ranks.

Next consider the distribution of the number of index entries in monographs, shown in Figure 3.3. The data is drawn from the Fondren Index Sample, which is described in detail in Chapter IV. Departure from linearity is clearly exhibited; this data is also shown in Figure 2.3, where it is plotted on lognormal probability paper with striking results which suggest that the points in Figure 3.3 should approximate a parabola. A portion of the parabola that fits this data is shown in the figure. The reader will notice certain peculiarities of the distribution of data points that are characteristic of this type of problem and lead to difficulties of estimation. First of all, small values of the independent variable--the number of index entries in this case---correspond to few data points if the data has been grouped for calculational convenience as these data have been. On the other hand, large values of the

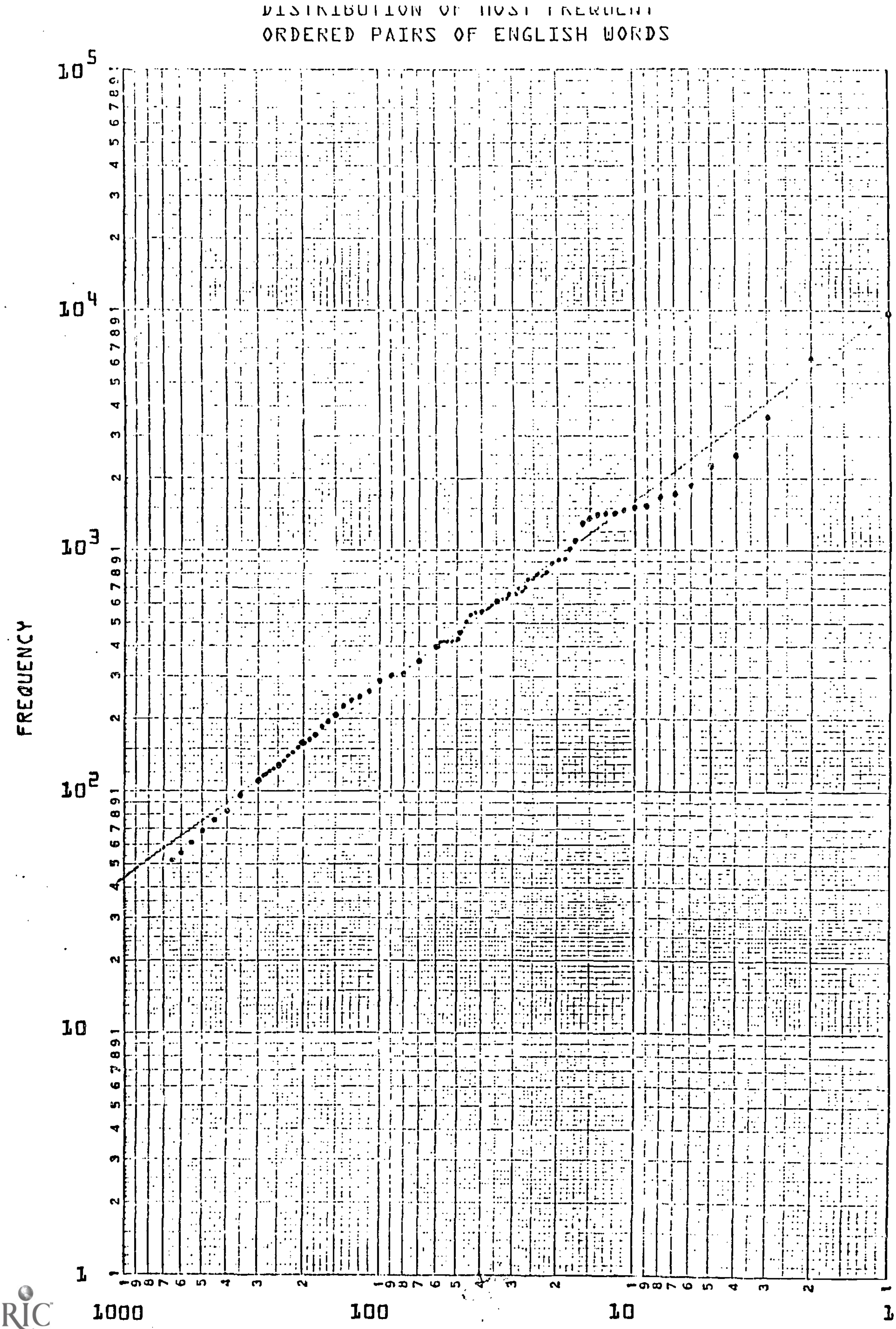
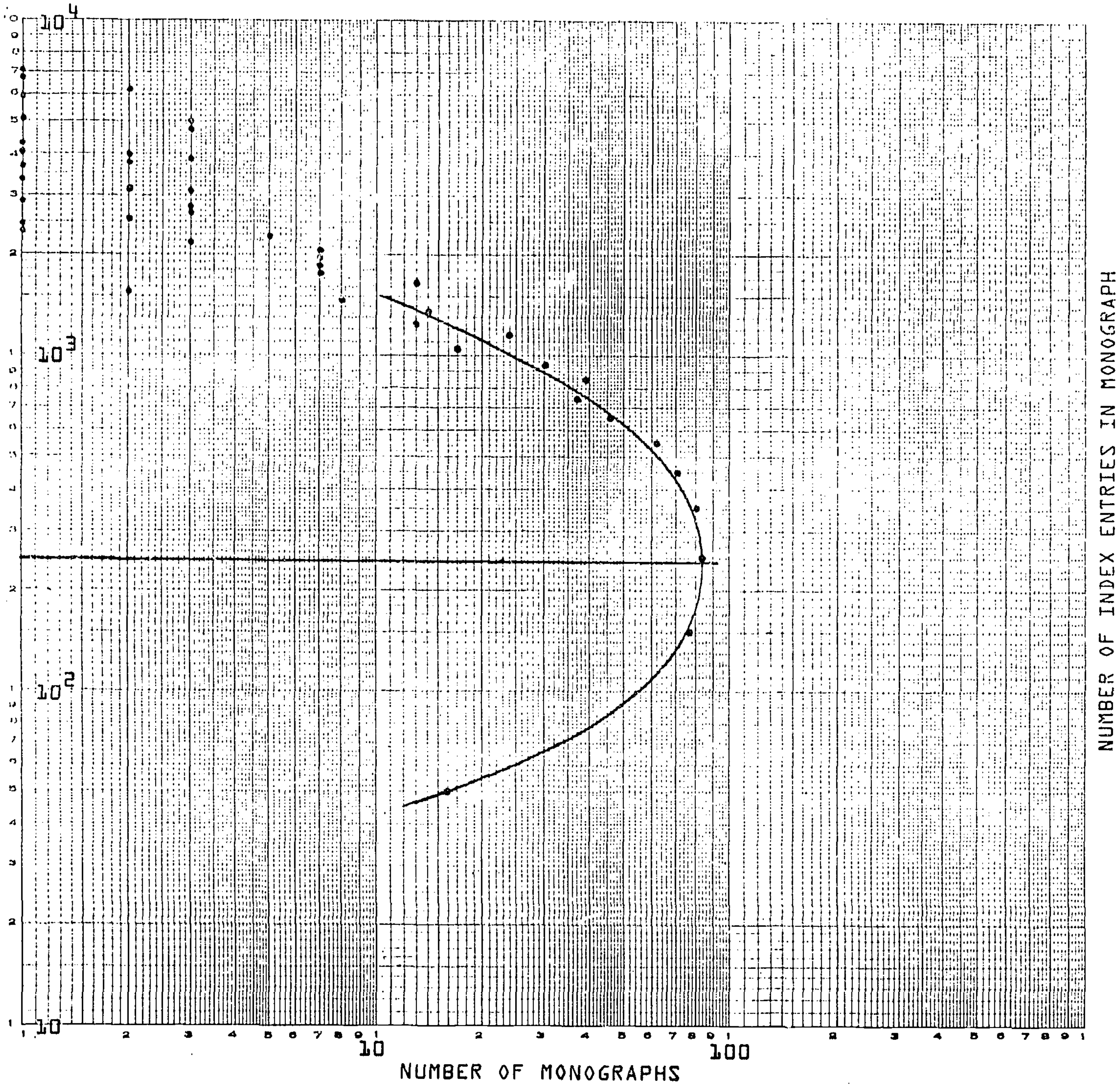

DISTRIBUTION OF MOST FREQUENT
ORDERED PAIRS OF ENGLISH WORDS
FREQUENCY
10^5
10^4
10^3
10^2
10
1
1000
100
10
1
RANK

Figure 3.2

Figure 3.3

INDEX ENTRY DISTRIBUTION
FROM
FONDREN INDEX SAMPLE



independent variable correspond to many data points even after grouping if the group intervals are of uniform size and for many of these the corresponding frequencies will coincide, leading to vertical 'segments' such as appear over the '1', '2', and '3' monograph markers. It is the geometric mean of these values that is important if the distribution is in fact lognormal.

More complex phenomena sometimes occur, and are perhaps most easily initially analyzed by studying the nature of the polynomial functions that approximate them, or at least portions of them, in the variables u and f(u). Consider, for example, the distribution of the number of pages in a monograph, shown in Figure 3.4. The data is drawn from the Fondren Sample (cp. Dolby, et. al. (19)) and has been grouped. From the figure one sees that monographs of fewer than 220 pages appear to follow the Zipf distribution whereas longer monographs have lengths that are well approximated by part of a lognormal distribution since they correspond to data points that fall nearly on part of a parabola as shown in the figure; the small arrows indicate computed values of points lying on the fitting parabola. We have no satisfactory explanation for this curious discontinuity in the effort function which describes the distribution of lengths of monographs which maximizes information per unit effort. Extraneous factors, perhaps related to the technology and economics of printing, are probably responsible but we have thus far been unable to isolate them. The reader should, however, compare Figures 2.5 and 2.6 which show the page distribution for unindexed and indexed monographs in the Fondren Sample.

Now consider the problem of determining the parameters of a lognormal distribution which represents given data. Take the equation of the lognormal in the form

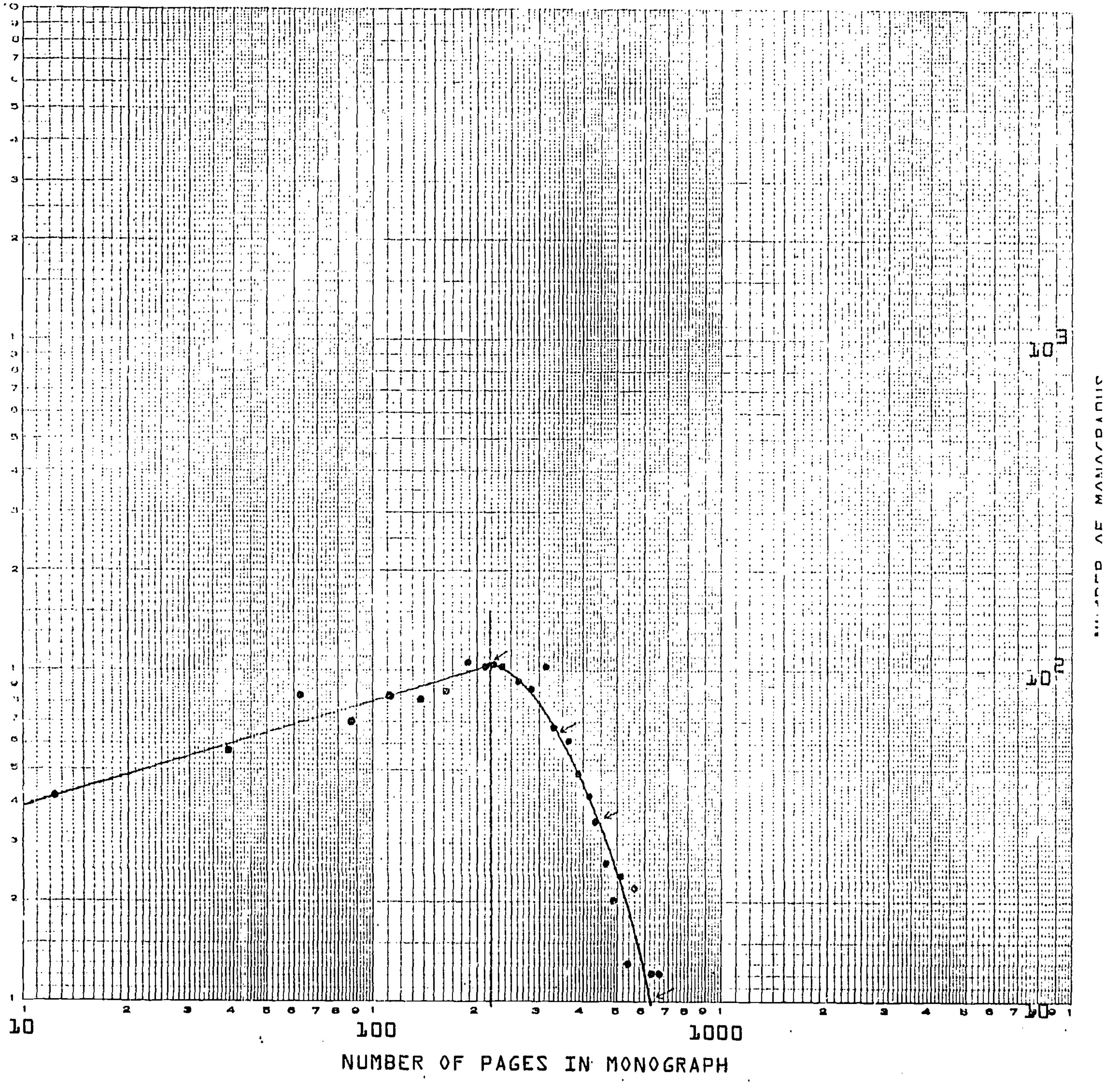
$$y= \frac{N}{s(x - a)\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\log(x-a) - m}{s}\right)^2} \qquad (3.22)$$

If the data consists of sample measurements $\{x_i\}$ such that the sample frequency of occurrence of x_i is y_i, then N is just the total number of measurements:

$$N = \Sigma y_i. \qquad (3.23)$$

Figure 3.4

PAGE DISTRIBUTION
FROM
FONDREN SAMPLE



The values of a, m, and s are determined by introducing an auxiliary quantity related to the skewness of the sample distribution, to whose definition we now turn.

Some more terminology is necessary. Define the k^th^ sample moment μ_k' by

$$\mu_k' = \frac{1}{N} \Sigma\, x_i{}^k y_i \ ;$$

μ_1' is the usual mean of the sample. If the sample moments are known, the central moments

$$\mu_k = \frac{1}{N} \ \Sigma (x_i - \mu_1')^k y_i$$

can be calculated. Expressions for the first four central moments in terms of the sample moments are seful when considering lognormal distributions. The first central moment μ_1 is evidently zero, since

$$\begin{aligned} \mu_1 &= \frac{1}{N} \Sigma \ (x_i - \mu_1') y_i \\ &= (1/N) \ \Sigma x_i y_i - (\mu_1'/N) \ \Sigma \ y_i \\ &= \mu_1' - \mu_1' \ . \end{aligned}$$

The next three central moments are given by the relations

$$\mu_2 = \mu_2' - (\mu_1')^2 \ ;$$

$$\mu_3 = \mu_3' - 3\mu_2'\mu_1' + 2(\mu_1')^3 \ ;$$

$$\mu_4 = \mu_4' - 4\mu_3'\mu_1' + 6\mu_2'(\mu_1')^2 - 3(\mu_1')^3 \ .$$

The positive square root of μ_2 is usually called the standard deviation and denoted by σ :

$$\sigma = \sqrt{\mu_2} \ .$$

Introduce the ratios

$$\beta_1 = \mu_3{}^2/\mu_2{}^3 \qquad (3.24)$$

and

$$\beta_2 = \mu_4/\mu_2{}^2 \ . \qquad (3.25)$$

β_1 is called the skewness of the sample; it provides a simple measure of the departure from symmetry about its mean. The skewness of the normal distribution, as for all symmetrical distributions, is 0. β_2 is sometimes known as the kurtosity of the sample distribution. If $\beta_2 < 3$, it is lower and flatter. The kurtosity of a normal distribution is 3.

In the literature, and unfortunately also in tables, other formulae are sometimes used to define quantities known as skewness and kurtosity. It is common, for instance, to find

$$\gamma_1 = \mu_3/\sigma^3 = \pm\sqrt{\beta_1}$$

called 'skewness' (note that the sign of γ_1 is the same as the sign of μ_3), and

$$\gamma_2 = \beta_2 - 3$$

is sometimes called 'kurtosity'. We follow Karl Pearson's usage, as found for instance in Ref. (20).

Skewness and kurtosity are of interest because they provide simple measures of the deviation of a sample distribution from the normal distribution and can be used to determine a family of distributions likely to provide an accurate and practical representation of data exhibiting skew variation; this procedure was introduced by Pearson (21). We will later have occasion to compare the lognormal representation of data occurring in information systems with representations by means of Pearson's distributions.

Skewness is of immediate interest here because the three numbers μ_1, μ_2, γ_1 determine the lognormal distribution that best fits given sample data. The unknown parameters a, m, and s of (3.22) can be expressed by means of an auxiliary quantity η which is the real root of the equation

$$\eta^3 + 3\eta - \gamma_1 = 0 \tag{3.27}$$

where γ_1 is the square root of the skewness (with the correct sign) as defined in (3.25). It is easy to see that there is in fact just one real root to (3.27), for otherwise there must be three, so the graph of the left side of (3.27) would have two turning points, which means that the derivative of the left side would have two real roots. But the derivative is $3\eta^2+3$, whose roots are pure imaginary.

The unique real root of the cubic equation (3.27) is readily and accurately approximated by using Newton's method. First select some reasonable approximation to the root; if a better choice is lacking, set

$$\eta_1 = \gamma_1/3$$

If η_k is the k^{th} approximation to the root η, then the next approximation is

$$\eta_{k+1} = \eta_k + (\gamma_1 - 3\eta_k - \eta_k^3)/(3\eta_k^2 + 3). \tag{3.28}$$

For example, if $\eta_1 = 3$, then $\eta_1 = 3/3 = 1$, and

$$\eta_2 = 1 + (3 - 3 - 1)/(3 + 3) = 0.833333\ldots ,$$

$$\eta_3 = 0.795556 ,$$

$$\eta_4 = 0.817973 ,$$

$$\eta_5 = 0.817732 = \eta_6 ;$$

therefore $\eta = 0.817732$ correct to six places.

Given a sufficiently accurate value of η, the parameters of the lognormal distribution (3.22) are (*cf*. Cramer, Ref. (22)):

$$a = \mu_1' - \sigma/\eta \ , \tag{3.29}$$

$$s = \{\log\ (1 + \eta^2)\ \}^{1/2}, \tag{3.30}$$

$$m = \log\ (\sigma/\eta) - \tfrac{1}{2}s^2 \quad . \tag{3.31}$$

Recall that $\sigma = \sqrt{\mu_2}$ is the standard deviation.

If the parameters of a lognormal distribution are known, it is of course possible to calculate the kurtosity of that distribution; the result is expressible in terms of η as

$$\beta_2 = \eta^8 + 6\eta^6 + 15\eta^4 + 16\eta^2 + 3 \ , \tag{3.32}$$

and is of no particular value except that it permits one to sketch the graph of the skewness and kurtosity pairs (β_1,β_2) that can belong to lognormal distributions. Figure 3.5 shows such a graph. The skewness and kurtosity of a particular data sample determine a point in the β_1-β_2 plane; the farther this point is from the lognormal curve, the less likely is the hypothesis that the sample data is drawn from a lognormal distribution.

By applying the usual techniques of the differential calculus it is easily shown that the maximum value of the lognormal distribution (3.22) is attained when

$$x = a + e^{m-s^2} \quad . \tag{3.33}$$

Some examples showing how these equations are to be applied to sample data may have some interest for the reader.

Consider first the distribution of the number of entries in a monograph index. For data drawn from the Fondren Index Sample, Figure 3.3 shows that this distribution apparently is lognormal. The following table, which summarizes the data in Table 4.3, groups the sample measurements according to intervals of 500 index entries and nominally associates them with the center value in each interval. This coarse grouping scheme

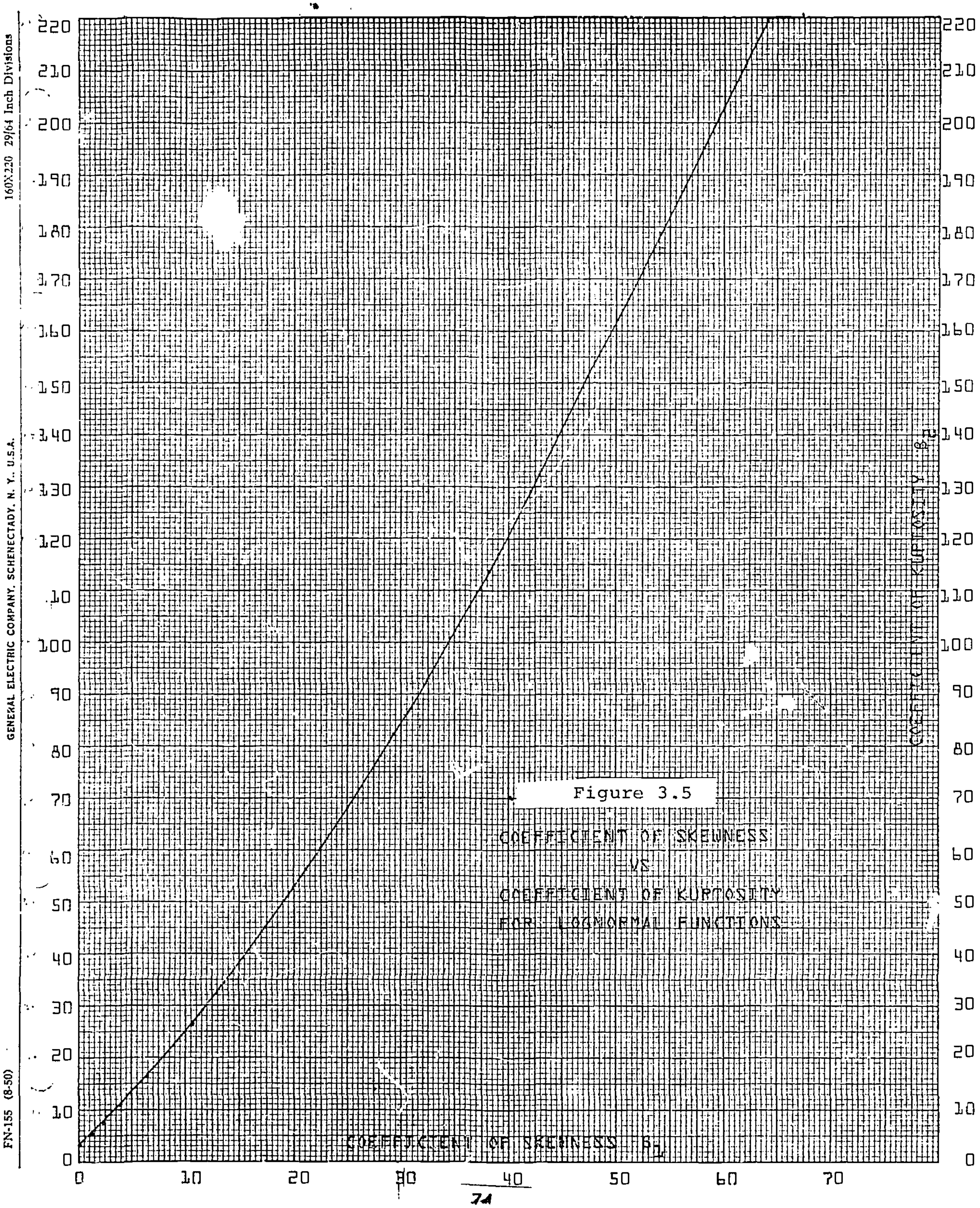


Figure 3.5

COEFFICIENT OF SKEWNESS
VS
COEFFICIENT OF KURTOSITY
FOR LOGNORMAL FUNCTIONS

decreases the number of categories to the point where the necessary calculations can be conveniently performed on a desktop calculator.

Table 3.1

GROUPED NUMBER OF INDEX ENTRIES FOR MONOGRAPHS

Nominal No. of Entries	Number of Indexes
250	326
750	214
1,250	76
1,750	36
2,250	17
2,750	9
3,250	6
3,750	8
4,250	2
4,750	6
5,250	1
5,750	1
6,250	2
6,750	1
7,250	1

One finds

$$N = 706$$

$$\mu_1' = 831.444759$$

$$\mu_2 = 878{,}281.765706, \quad \sigma = 937.166882$$

$$\mu_3 = 2{,}551{,}331{,}421.74$$

$$\mu_4 = 13{,}238{,}646{,}211{,}200$$

to twelve figures. Consequently

$$\gamma_1 = 3.0997$$

$$\beta_1 = 9.6080$$

$$\beta_2 = 17.1623 \ .$$

According to Figure 3.5, the sample value of β_1 corresponds to $\beta_2 \cong 24$ if the distribution is lognormal. The disagreement is not serious in view of the effect of grouping the data and the small size of the sample; regarding the latter point, J. Carroll's remarks in Reference (16) are instructive. For the original ungrouped data, it turns out that $\mu_1' \cong 836$, so the effect of grouping the data is not entirely negligible.

These estimates imply $\eta = 0.837450$ correct to six figures; indeed, since γ_1 is nearly equal to 3, it is a good idea to select the solution to (3.27) previously calculated as an example as the starting value η_1 of the approximation procedure, leading to

$$\eta_1 = 0.817732,$$

$$\eta_2 = 0.837642,$$

$$\eta_3 = 0.837450 = \eta_4 = \eta \ .$$

Substitution in (3.29)-(3.31) produces the parameters of the lognormal:

$$a = -287.6 \ ,$$

$$s = 0.7284 \ ,$$

$$m = 6.755 \ ,$$

so

$$y = \frac{706}{0.7284(x+287.6)\sqrt{2\pi}} \; e^{-\frac{1}{2}\left\{\frac{\log(x+287.6)-6.755}{0.7284}\right\}^2}$$

Observe that

$$c^m \cong 858 \cong 874 \cong K^2.$$

According to these calculations, the (grouped) number of index entries behaves as if approximately 288 entries are 'missing' from indexes. To what extent this must be attributed to the effect of grouping the data so coarsely we have not attempted to determine except to note that the intervals that were chosen will tend to produce this type of qualitative effect because most of the indexes represented in the first group (those having fewer than 500 entries) contain more than the 250 entries nominally ascribed to that category, thus biassing the distribution toward low values.

The next example is a particularly useful pedagogical illustration because the data are unusually regular and occur in large number in a form convenient for calculations, but it is of considerable independent interest as well. Consider the distribution of dictionary words according to the number of vowel strings they contain. The number of vowel strings, in the technical sense in which it is used here, is a graphemic substitute for the phonemic notion of the number of syllables contained in the spoken form of a word; the precise definition we use is given in Reference (23), but will not be necessary for our present purposes since the intuitive correspondence with the notion of syllable is sufficiently accurate. The words under consideration are the 64,041 lexed words of the Shorter Oxford Dictionary which contain at least one vowel string. Figure 3.6 displays the data on bi-logarithmic graph paper; the general parabolic tendency is apparent. From the data given in Table 3.2 one readily calculates

$$N = 64{,}041 \ ,$$

$$\mu_1 = 2.6889 \ ,$$

$$\mu_2 = 1.1096 \ , \ \sigma = 1.0534 \ ,$$

$$\mu_3 = 0.5027 \ ,$$

$$\mu_4 = 3.7011 \ ,$$

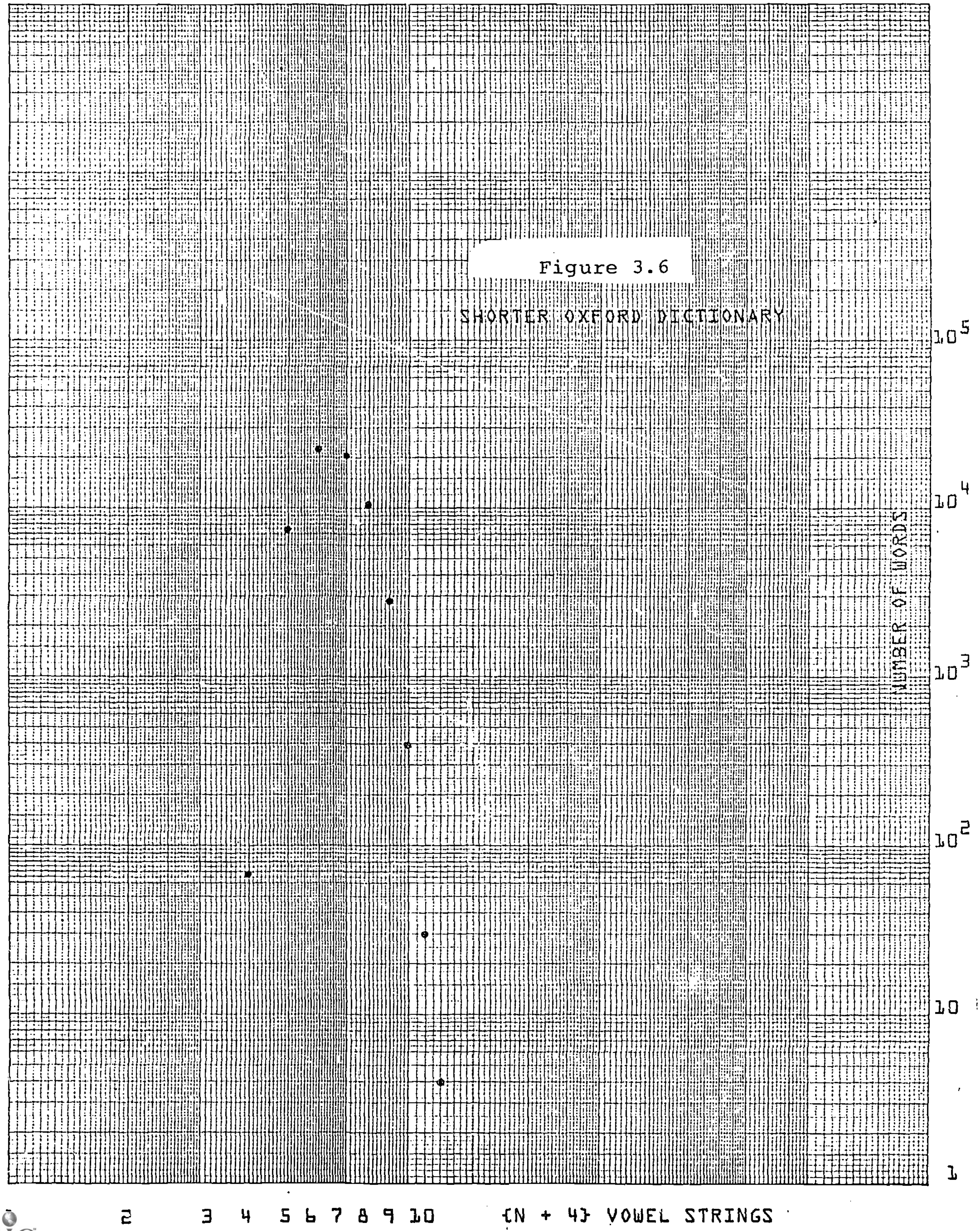


Figure 3.6

SHORTER OXFORD DICTIONARY

TABLE 3.2

DISTRIBUTION OF LEXED WORDS
FROM THE SHORTER OXFORD DICTIONARY
BY NUMBER OF VOWEL STRINGS

Number of Vowel Strings	Observed Number of Words	Calculated Number of Words
0	63	285
1	7,158	6,618
2	22,568	22,160
3	20,762	22,072
4	10,293	9,737
5	2,770	2,531
6	393	691
7	30	178
8	4	24

So

$$\gamma_1 = 0.4301 ,$$
$$\beta_1 = 0.1850 ,$$
$$\beta_2 = 3{,}0060 .$$

For a lognormal distribution, this value of skewness implies a kurtosity approximately equal to 3.3, which agrees reasonably well with the value computed from the sample. The calculated parameters of the lognormal fitting these data are

$$\begin{aligned} a &= 4.7086 , \\ s &= 0.1407 , \\ m &= 1.9916 . \end{aligned} \qquad (3.34)$$

The maximum value of this lognormal distribution is $y \cong 25{,}000$ and occurs at $x \cong 2.46$. The rightmost column of Table 3.2 shows the number of words as a function of the number of vowel strings as calculated from the distribution defined by the parameters given in (3.34) above.

Some degree of caution must be exercised when one attempts to determine if data can be reasonably fitted by a lognormal function. If the entire range of the variable is not represented by the data due either to unfortunate grouping or absence of information, graphical representation of the data may be misleading. We will describe one way in which this can happen. Suppose that $\{(x,y)\}$ is a data sample exhibiting the frequency function of some variable x. Let $\{(x,Y)\}$ denote the cumulative frequency distribution defined by

$$Y(x) = \text{sum of } y(x_0) \text{ for } x_0 \geq x .$$

For instance, if y(x) denotes the number of individuals having an annual income of x dollars (more exactly, x+b dollars for some conveniently chosen small increment b], then Y(x) denotes the number of individuals with income at least x dollars. The latter cumulative distribution is exact in the sense that it presents the actual number of individuals belonging to the corresponding category of the sample, whereas the frequency distribution presents grouped data as a substitute for frequency density functions and therefore potentially introduces error into the sample data. For this reason it is often desirable to analyze cumulative sample distributions rather than the corresponding approximate frequency functions

Consider, therefore, a cumulative distribution $\{(x,Y)\}$ whose points fall close to a straight line when exhibited on log-log graph paper. In this event it is reasonable to conclude that log Y = a log x + log c so

$$Y = cx^a : \qquad (3.35)$$

Y is a power function of x. The frequency function can be retrieved from (3.35) from the relation

$$y = dY/dx ;$$

we find

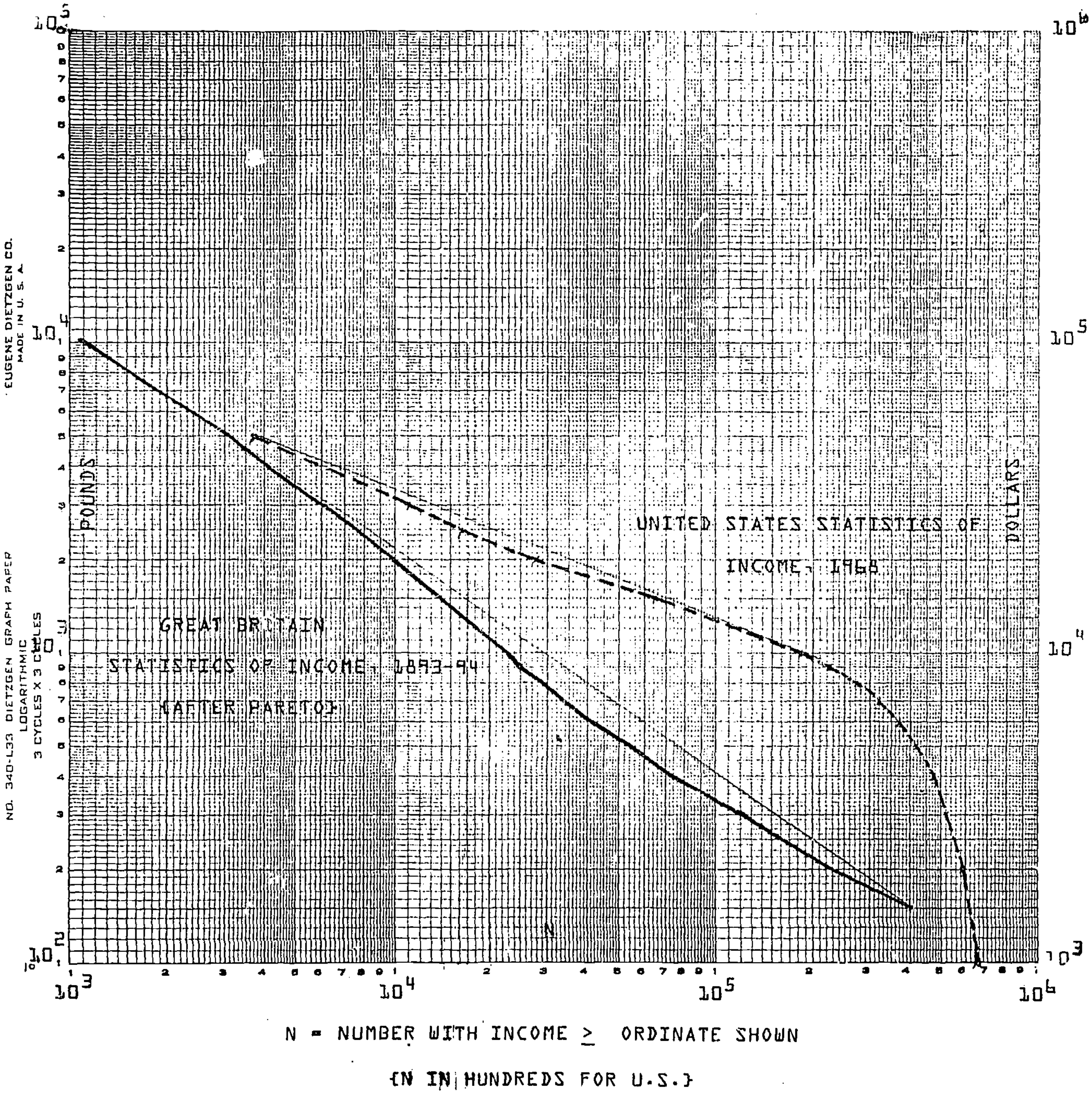
$$y = cax^{a-1} ,$$

so y is also a power function of x.

If the entire range of the variable x is not represented by the sample data, this procedure of determining the theoretical frequency function from its cumulative distribution by differentiation can be misleading. Figure 3.7

Figure 3.7

DISTRIBUTION OF INCOME



displays the cumulative income distribution for Great Britain in 1893-94 and for the United States in 1968 drawn on log-log graph paper. The data for the former fall along a straight line which implies that it and hence also the associated frequency function are power functions. This is the famous 'law' of Vilfredo Pareto (Ref. (6), vol.2, p. 304 et seq.). The leftmost part of the corresponding distribution for the United States also exhibits a generally linear trend but the rightmost portion cannot be so construed at all. Several interpretations of this anomoly are possible, including some that are based on variations between the underlying economic and social structures of the two nations during the two time periods surveyed, but it is possible to account for the apparent contradiction by examining the extent to which the sample data represents the entire range of variation.

The United States data refers to income of any size reported on tax returns, of which more than 73 million were tallied. The data used by Pareto refers only to incomes greater than 150 pounds sterling per year, of which 400,648 were reported. The number of inhabitants per income reported was nearly 3 for the United States in 1968; using this figure, we see that if Pareto's data includes essentially all incomes, the population of Great Britain in 1893-94 should have been about 1.2 million. It was in fact perhaps greater than 8 million, which suggests that there were possibly more than two million people in Great Britain in those years having a positive income less than 150 pounds per year. The reader should not interpret this estimate as anything but a very crude indication of the number of incomes that were probably overlooked in the data sample. Now apply this estimate to extend the graph for Great Britain in Figure 3.7; the extension must turn downward when the total number of incomes exceeds about 2.4 million, so the extended income curve will have the same general appearance as that for the United States.

As has already been remarked, the cumulative distribution of income function for the United States certainly is not a power function and therefore the corresponding frequency function cannot be either. By plotting the grouped frequency data published by the Internal Revenue Service on lognormal probability graph paper, it can be seen that the frequency function of income distribution can be approximated throughout its entire range by a lognormal function with the parameter a of eq (3.22) approximately equal to $4,000. It is likely that the income distribution for Great Britain used by Pareto can also be approximated by a lognormal function, but it is necessary to have an accurate estimate of the total number of incomes less than 150 pounds per year before one can calculate the cumulative fractions necessary for the employment of lognormal probability graph paper or the methods for estimating the parameter values from sample data given earlier in this Chapter.

References

1. Estoup, J. B., *Gammes stenographiques*, 4th Edition, 1916.

2. Bradford, S. C., "Sources of Information on Specific Subjects", *Engineering*, (1934), January 26.

3. Mandelbrot, B. "An Information Theory of the Statistical Structure of Language", *Proceedings of the Symposium on Applications of Communication Theory*, Butterworth, 1953.

4. Mandelbrot, B., "On the Language of Txonomy: an Outline of a 'Thermostatistical' theory of Systems of Categories with Willis (natural) Structure", *Information Theory; Papers Read at a Symposium on Information Theory, London 1955*, Butterworth, 1956, 135-45.

5. Zipt, G. K., *Human Behavior and the Principle of Least Effort*, Addison Wesley, 1949.

6. Pareto, V., *Cours d'economie, politique*, Lausanne, 1897.

7. Lotka, A. J., "The Frequency Distribution of Scientific Productivity", *Journal of the Washington Academy of Sciences*, 16(1926), 317.

8. Price, D. J. de solla, *Little Science, Big Science*, Columbia University Press, 1963.

9. Schrodinger, E., *Statistical Thermodynamics*, Cambridge University Press, Cambridge, 1964.

10. Yule, G. U., The Statistical Study of Literary Vocabulary, Cambridge: The University Press, 1944.

11. Belonogov, G. G., "On some Statistical Regularities in Written Russian", Vopr. Jazykoznanija, 7(1962), 100, (in Russian).

12. Good, I. J., "Statistics of Language: Introduction", Encyclopaedia of Linguistics, Information, and Control, Pergamon Press, London, 1969, 567-81.

13. Houston, N. and E. Wall, "The Distribution of Term Usage in Manipulative Indexes", American Documentation, 15(1964), 105-14.

14. Fairthorne, R. A., "Empirical Hyperbolic Distributions (Bradford-Zipf-Mandelbrot) for Bibliometric Description and Prediction", Journal of Documentation, 25(1969) 319-43.

15. Wall, E., "Further Implications of the Distribution of Index Term Usage", Parameters of Information science: Proceedings of the American Documentation Institute Annual Meeting, 1964, Volume 1, American Documentation Institute, 1964, 457-66.

16. Carroll, J. B., "On Sampling from a Lognormal Model of Word Frequency Distribution", Computational Analysis of Present-Day American English (Henry Kucera and W. Nelson Francis), Brown University Press, Providence, Rhode Island, 1967, 406-24.

17. Kucera, H., and W. N. Francis, Computational Analysis of Present-Day American English, Brown University Press, Providence, Rhode Island, 1967.

18. Shannon, C. E., "Prediction and Entropy of Printed English", Bell System Technical Journal, 30(1951), 50.

19. Dolby, J. L., V. J. Forsyth, and H. L. Resnikoff, Computerized Library Catalogs: Their Growth, Cost, and Utility, M.I.T. Press, Cambridge, 1969.

20. Pearson, E. S., and Hartley, H. O., Biometrika Tables for Statisticans, Volume I., Cambridge, England: The University Press, 1954.

21. Pearson, Karl, "Mathematical Contributions to the Theory of Evolution", Philosophical Transactions, A, 186(1895), 343-414; 197(1901), 443-59; 216(1916), 429-57.

22. Cramer, H., *Mathematical Methods of Statistics*, Princeton University Press, Princeton, 1946.

23. Dolby, J. L., and H. L. Resnikoff, "On the Structure of Written English Words", Language, 40(1964), 167-96.

CHAPTER IV

THE STRUCTURE OF BACK OF THE BOOK INDEXES

THE STRUCTURE OF BACK OF THE BOOK INDEXES

Book indexes are among the most common and most ancient access mechanisms, although they have not always been loved. Glanville, in Vanity of Dogmatizing, said:

> Methinks 'tis a pitiful piece of knowledge that can be learnt from an index, and a poor ambition to be rich in the inventory of another's treasure,

and more recently T. E. Lawrence wrote:

> ...half-way through the labor of an index to this book I recalled the practice of my ten years' study of history; and realized I had never used the index of a book fit to read.

However, as an unnamed contributor to a recent edition of the Encyclopedia Britannica put it,

> (It has) become almost a sine qua non that any good book must have its own index.

Indeed, as we shall see below, more than one-third of all non-serial items in the shelf list of a medium size university library do contain an index, and it seems as if the back of the book index is not only here to stay but is in the process of spawning a genus of related tools for indicating "the position of information on any given subject".

The object of this chapter is to study indexes to books in order to determine what structure, if any, they possess. It is not surprising that indexes* exhibit great variability in size, content, and utility, which makes it difficult to assess their nature in general from an examination of one or several exemplars. We have elected to study indexes in three ways.

*Throughout this chapter 'index' will only refer to back-of-the-book indexes.

The first and most reliable way is based on the selection of a random sample of book indexes. Such a sample has been assembled by extraction of the indexes from all monographs represented in a random sample of the shelf list of a medium size university library; it consists of approximately six hundred thousand index terms spread throughout some 700 books, and will be described in what follows:

The second means of studying indexes is concerned with the structure exhibited by each index separately. Information of this sort cannot be obtained from statistical agglomerations; rather it demands that indexes be considered in detail and the resulting structures, if any are found, compared for a sample of indexes.

A book index directs the user to the location of specified information in the book to which it refers. Should the book in question not contain any indexed information about the subject of interest, the inquirer is left to continue his search in the indexes of other unspecified books. There are, of course, several indirect methods for deciding how the next book in the search process should be selected, utilizing information contained in the bibliographies or the linear shelf list order determined by a subject classification scheme such as that of the Library of Congress, but none of these have the virtue of immediacy nor of completeness. Our third means of studying indexes is based on a cumulative index to 80 books in the field of statistics. It appears to offer attractive efficiencies in the information search process while it provides a view of the overall structure of the field itself.

The Fondren Index Sample is a random sample of 668 monograph shelf list cards corresponding to indexed books. Multiple volumes catalogued on one shelf list card increase the sample somewhat so that a total of 706 indexes are represented.

The Fondren Index Sample is a subsample of the Fondren Sample, which is a random sample of cards drawn from the shelf list of the Fondren Library at Rice University. The Fondren Sample is described in some detail in Reference [1]. Analyses of the sample may be expected to accurately reflect the structure of library collections to the extent that they are similar to the Fondren collection; in particular, the archival collections of medium size university libraries are probably generally similar although certain special fields may be more or less well represented. For instance, the Fondren collection is particularly weak in law, medicine, and Russian language and literature, and strong in chemistry. These differences are unlikely to play a significant role in determining the reliability of the sample for studying index structure since indexes are relatively insensitive to the nature of the subject material to which they refer; the gross category differences, as between science and fine arts, are, as will be shown below, substantial, but the Fondren collection encompasses adequate representation in each of such broad categories.

There are special problems associated with the analysis of complex data drawn from any sampling process. The index sample is no exception. Some of the sample indexes have a format so unusual as to make them incomparable with the average index; a small number were written in non-Roman alphabets so we were unable to correctly identify the structural features of interest. Because the fraction of anomolous indexes was small, it was decided to delete them from the index sample for this initial study.

This decision was bolstered by another complication; not all of the books represented by the original random sample could be located for the present study, which took place about two years after the original selection of shelf list cards. The number of unlocatable items was 33, approximately 1.7% of the Fondren Sample; this is the effective rate of loss for the two year period in the sense that the usual mechanisms for tracking items not present on the shelf in their proper location were applied without success for these items, noting that just prior to the selection of the sample the shelf list had been checked against the shelf and weeded. This suggests that slightly less than 1% of the monograph archive is lost each year.

If all 33 unlocatable items had had indexes, they would have constituted nearly 4% of the index sample; items excluded for special reasons such as language or format incompatibility totalled 22. Therefore, not more than 7.5% and more likely not more than 4.5% of the indexed volumes in the Fondren Sample have been excluded from the index sample. With this preliminary in mind we can now turn to the consideration of the index sample.

First observe that not all monographs are candidates for indexing; we have found no Library of Congress class "A" items in the sample which contain an index, and therefore class "A" is excluded from all further considerations. Similarly, neither maps nor musical scores are indexible in the "back of the book" sense, so they too are excluded. Excluding these items and all serial publications, one finds that there are 1,830 relevant items in the Fondren sample. Of these, 668 have indexes; thus we find that 37% of the monographs in the Fondren sample contain indexes.

As previously noted, the 668 LC cards lead to a total of 706 volumes with indexes. The distribution of these 706 volumes by LC class is shown in Table 4.1 together with the fraction that is indexed for each class. This fraction runs from a low of 0.18 for N (Fine Arts) and P (Language) to a high of 0.61 for Q (Science) and 0.67 for Naval Science.

Table 4.1 also provides the mean number of index entries per book indexed. The grand mean for the collection is 836 index entries per book, with the class means varying from a high of 1,391 entries per book for class F (U.S. Local History) to a low of 614 for class J (Political Science).

The product of these two measures provides an average measure of the amount of access per book in the collection and in each of its subsets. This distribution is shown separately in Table 4.2. This list breaks rather naturally into three subsets of nearly the same size. The first seven categories (classes F, G, V, K, D, E, and Q) would seem to share the property that they are all primarily concerned with careful description of the world as it is and as it has been. The middle group (classes H, C, R, T, Z, L, and J) is primarily devoted to man's effort to cope with the environment described so carefully in the first group. The lowest group appears a bit anomolous in that it contains the core of the arts: music, philosophy, religion, language, literature, and the fine arts as well as the more mundane but ever present categories of war and agriculture. Although we should not like to make too much of this particular arrangement of the LC classes, Table 4.2 does provide an interesting example of the insight one gains into the use of the system of literary stores by rather elementary counting procedures.

The index sample consists of a total of 590,329 index entries spread across the 706 indexes. Table 4.3 lists the number of indexes as a function of the number of entries they contain, grouped by hundreds of index entries. Figure 4.1 exhibits the lognormality by showing the data of Table 4.3 plotted on lognormal paper. The standard deviation on the log scale is 0.442 which is at the upper end of the range for log-length distributions given in Chapter II.

Table 4.1

FONDREN SAMPLE: FRACTION OF SAMPLE ITEMS CONTAINING AN INDEX, BY LC LETTER CLASS

Class	Mean Number of Entries per Index	Fraction Indexed (rounded)	Fraction Class is of Fondren Sample	Short Class Name
B	667	.31	.100	Philosophy-Religion
C	690	.53	.009	History-Auxiliary Sciences
D	1,102	.51	.095	History & Topography (except America)
E	1,062	.49	.040	American (General) & U.S. (General)
F	1,391	.46	.027	United States (Local) & America (ex. U.S.)
G	1,264	.50	.011	Geography-Anthropology
H	697	.54	.104	Social Sciences
J	614	.46	.023	Political Science
K	1,375	.43	.004	Law
L	620	.49	.038	Education
M	915	.25	.015	Music
N	615	.18	.033	Fine Arts
P	714	.18	.300	Language & Literature
Q	850	.61	.093	Science
R	716	.50	.010	Medicine
S	638	.20	.006	Agriculture-Plant & Animal Husbandry
T	707	.47	.032	Technology
U	840	.22	.010	Military Science
V	934	.67	.005	Naval Science
Z	1,328	.24	.023	Bibliography & Library Science

Total relevant items in Fondren Sample = 1823

Number of these items indexed = 668

Fraction indexed = 668/1830 = 0.37

Table 4.2

INDEX ACCESS BY LC CLASS

LC Class	Mean No. Index Entries per Book	Short Class Name
F	640	U. S. Local History
G	632	Geography
V	626	Naval Science
K	591	Law
D	562	World History
E	520	U. S. History
Q	519	Science
H	376	Social Science
C	366	Auxiliary Sciences (History)
R	358	Medicine
T	332	Technology
Z	319	Library Science
L	304	Education
J	282	Political Science
M	229	Music
B	207	Philosophy-Religion
V	185	Military Science
P	129	Language Literature
S	128	Agriculture
N	111	Fine Arts

Table 4.3

FREQUENCY OF INDEX ENTRIES FOR ITEMS
IN THE FONDREN INDEX SAMPLE

Number of Index Entries	Number of Indexes	Cumulative Number of Indexes	Cumulative Fraction of Indexes
0 - 99	16	16	.023
100 - 199	77	93	.132
200 - 299	83	176	.249
300 - 399	80	256	.362
400 - 499	70	326	.462
500 - 599	62	388	.549
600 - 699	46	434	.615
700 - 799	37	471	.667
800 - 899	39	510	.722
900 - 999	30	540	.765
1000 - 1099	17	557	.789
1100 - 1199	24	581	.823
1200 - 1299	13	594	.841
1300 - 1399	14	608	.861
1400 - 1499	8	616	.872
1500 - 1599	2	618	.875
1600 - 1699	13	631	.893
1700 - 1799	7	638	.903
1800 - 1899	7	645	.913
1900 - 1999	7	652	.923
2000 - 2099	7	659	.933
2100 - 2199	3	662	.937
2200 - 2299	5	667	.944
2300 - 2399	1	668	.946
2400 - 2499	1	669	.947

Table 4.3
(Continued)

2500 - 2599	2	671	.950
2600 - 2699	3	674	.955
2700 - 2799	3	677	.959
2800 - 2899	1	678	.960

3000 - 3099	3	681	.964
3100 - 3199	2	683	.967
3300 - 3399	1	684	.969
3500 - 3599	1	685	.970
3700 - 3799	2	687	.973
3800 - 3899	3	690	.977
3900 - 3999	2	692	.980
4000 - 4099	1	693	.981
4200 - 4299	1	694	.983
4700 - 4799	3	697	.987
4900 - 4999	3	700	.991
5100 - 5199	1	701	.993
5900 - 5999	1	702	.994
6200 - 6299	2	704	.997
6700 - 6799	1	705	.998
7000 - 7099	1	706	1.000

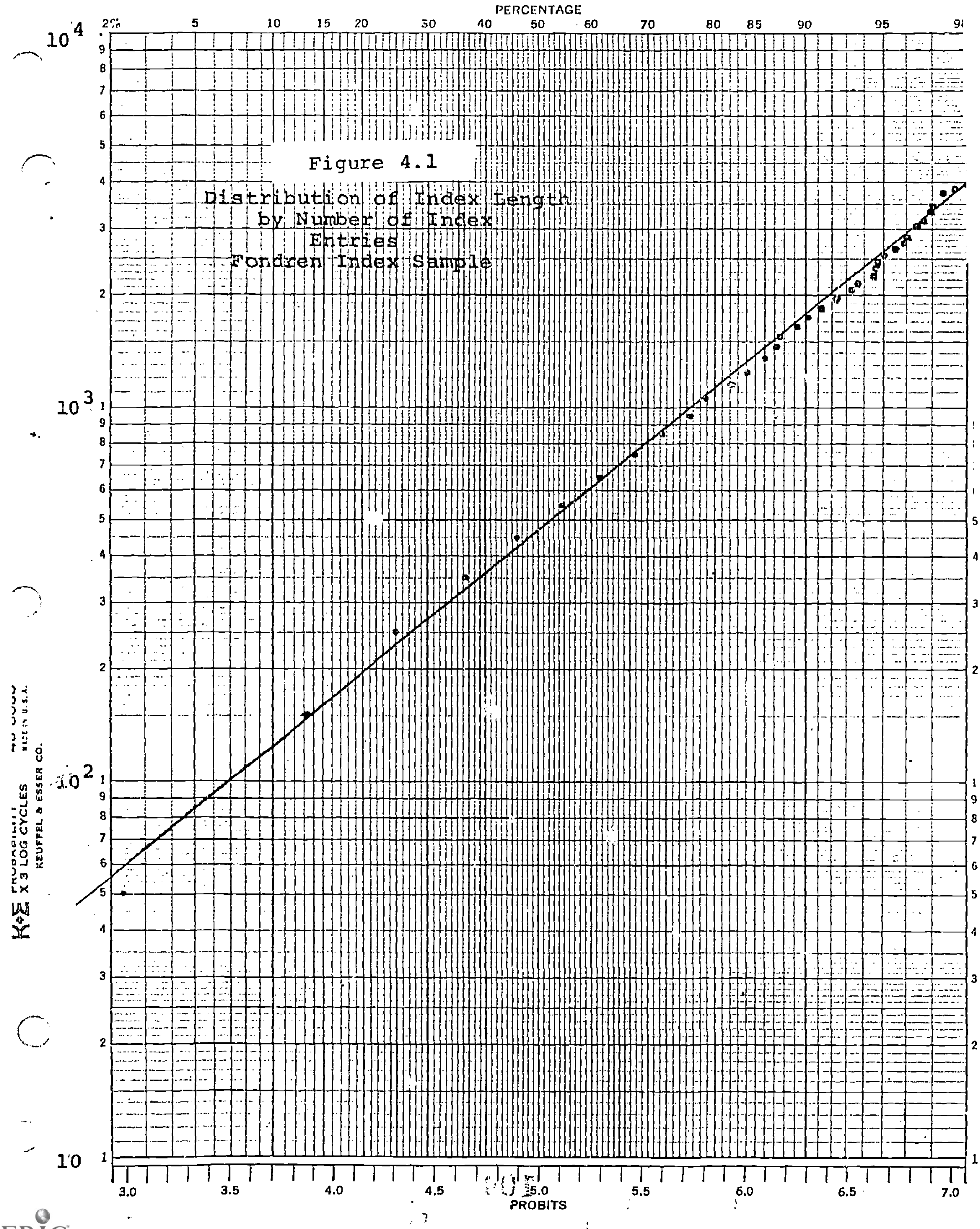


Figure 4.1

Distribution of Index Length by Number of Index Entries Fondren Index Sample

A distinction should be made between the number of index entries in an index and the number of locations to which these entries refer. The former quantity is the number of distinct word sequences appearing in an index, and is an absolute measure of index size which is independent of the details of format and page composition; the latter is usually the number of page locations referred to in an index, which clearly depends on the size of the page. In the Fondren sample of indexed books there are, on the average, 1.8 page locations per index entry. Thus, the 836 (average) distinct entries refer, on the average to 1,505 text locations. As there are on the average 341.5 pages per indexed book, there are 4.4 indexed text locations per page. Roughly speaking, this means that there is one index page location for each five sentences of text.

The aggregate size of the index as printed can be determined by estimating the average number of characters per entry and multiplying by the average number of entries. A preliminary estimate of the average number of characters was obtained by counting the entries in the cumulative index to statistical books (discussed at greater length in Chapter VI) as the format of the material is in particularly nice form for counting purposes. This estimate shows that the entries are about 25.47 characters in length exclusive of page location information. If, as in Chapter I, this is augmented by 4 characters per entry to include the typical page location reference information, then the average index of 836 entries consists of 24,637 characters and therefore the ratio of indexed book size to index size is about 33.27 to 1.

These global statistics provide a direct measure of the proportion of the monograph collection that is devoted to what might be called "self access". The aggreement of the access ratio (of about 30 to 1) with other access ratios developed in Chapter II helps to solidify the foundations of the level structured access model. Given the difficulty of assessing the quality of indexing (see (2) and the references therein) these statistics also provide the foundation of a basis for comparing various indexing procedures, particularly for comparing algorithmically derived indexes to manual indexes. The fundamental regularities of the length measures discussed here suggest that an algorithmically prepared index must at least be of the correct overall size to be of any use at all.

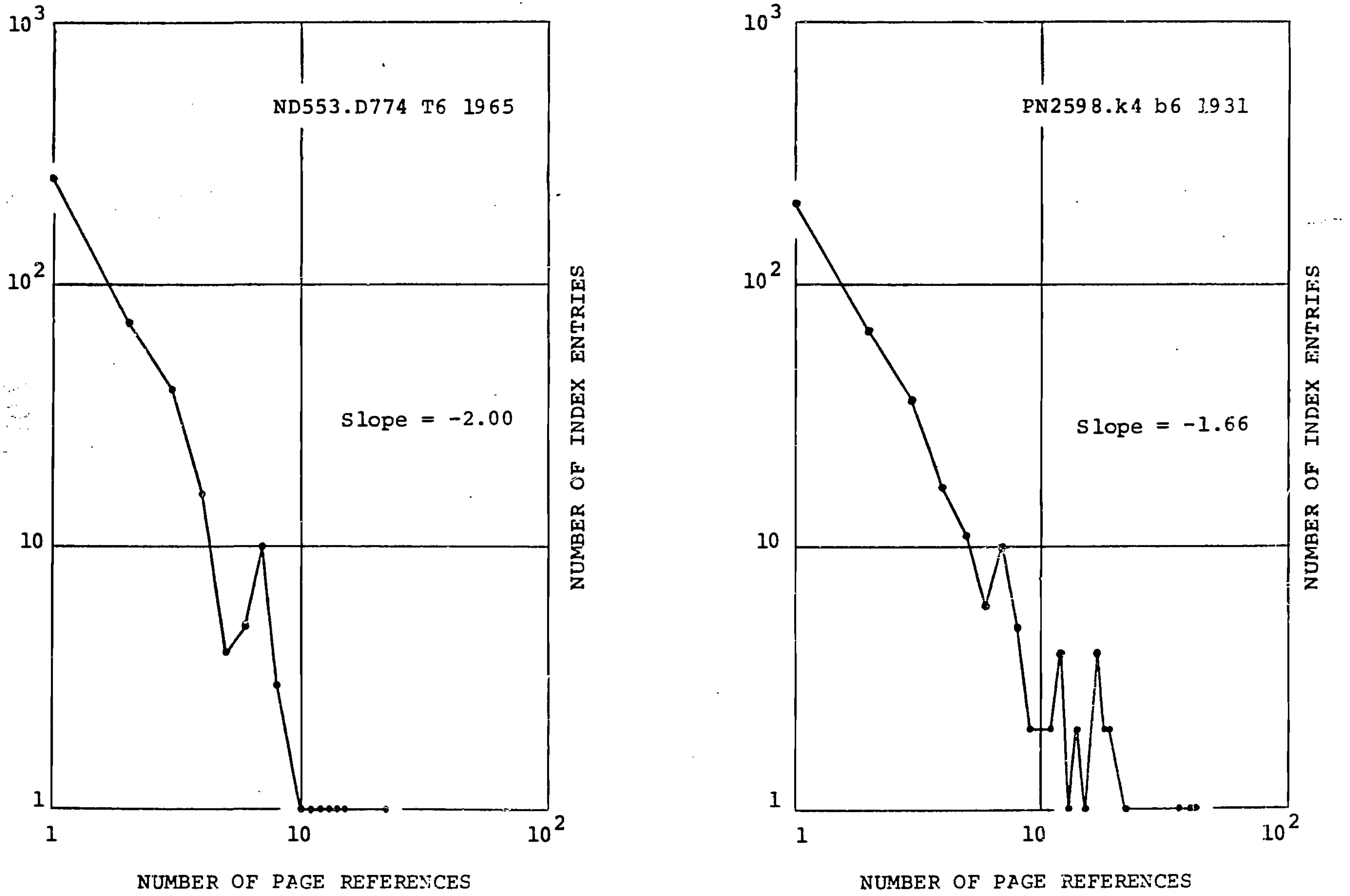
The find structure of the individual indexes can presumably shed more light on the situation. For these purposes, we have selected a random sub-sample of 28 indexes from the main Fondren Index Sample. For each of these indexes we have determined the distribution of the number of entries with one, two, three...page locations per entry. This distribution is comparable to the "frequency of frequencies" problem discussed extensively by Zipf, Bradford, Mandelbrot, et al (see Chapter III). Were the index an extractive index (i.e., one that is derived by extracting sequences of words from the text and inserting these sequences without change in the index) and were the page locations explicitly tied to the position on the page so that multiple occurrences of the entry on a single page would occur multiply in the index, then it might be anticipated that the text location distribution of index entries would be Zipf-Mandelbrot distribution which would arise from the phrases which are the index entries in the same way as the usual Zipf distribution arises from text word occurrences.

However, indexing practice normally requires a set of sophisticated transformations from the running text to the index and also reduces multiple entries on a page to a single page location. Further, not all "phrases" are indexed and it would appear that those which are left out are among both the most frequently occurring and least frequently occurring. Nevertheless, it seems reasonable to approach the problem at the first order of approximation by assuming a model of the Zipf-Bradford-Mandelbrot type; i.e., by examining the form of the distribution on log-log graph paper. This has been done for all 35 of the sample indexes, all 28 of which are presented here (Figures 4.2). (The remaining graphs appear in Appendix II.) The plots are given in the converse form to that used by Zipf in order to provide the converse form to that used by Zipf in order to provide stability (see Kendall (3)). Thus the largest point on the graph represents the number of index entries with single page references rather than the number of page references for the most frequently referenced item.

Two graphs shown are typical for the sample as a whole. In almost every case a straight line provides a reasonable approximation, with slopes ranging from roughly -1.1 to 05.5. Thus the Zipf-Mandelbrot approximation holds well for index location frequency distributions. The importance of the slope as a parameter of index measurement can be seen by recalling the Mandelbrot formulation which maximized the expected information per unit effort; the reader may find it useful to compare e.g. (3.5) ff:

Figure 4.2

Number of Index Entries vs. Number of Page References



$$I = \frac{-\Sigma\, p(x) \log n}{\Sigma\, p(x) \log x} \qquad (4.1)$$

The function that maximizes this ratio is the Zipf-Mandelbrot distribution:

$$p(x) = c\, x^{-s} \qquad (4.2)$$

Substitution of (4.2) into (4.1) yields

$$I = \frac{-\Sigma\,(\log c - s \log x)\, cx^{-s}}{\Sigma\, cx^{-s} \log x} = \frac{\Sigma x^{-s} \log \Sigma x^{-s}}{\Sigma x^{-s} \log x} \qquad (4.3)$$

where all logarithms are to the base e and the summations extend from 1 to the maximum number of page references per index entries.

For s greater than one, the summations all converge to functions of the Riemann zeta function as the maximum number of page references per index entry increases. Hence, with the sums running overall positive integers,

$$I = s - \frac{\zeta(s) \log \zeta(s)}{\zeta'(s)} \qquad (4.4)$$

An s increases, the ratio on the right, in turn, converges to $(\log 2)^{-1} = 1.443$ so that a first order approximation to Mandelbrot information for the Zipf-Mandelbrot form is given by

$$I = s + 1.443$$

For s greater than or equal to 3, the error is less than 10%. In other words, to a first order approximation, Mandelbrot's measure of information per effort is directly proportional to s, the negative value of the slope of the approximating straight line on log-log paper.

For data that perfectly fits the Zipf-Mandelbrot model, the parameter s can be determined from the relation:

$$s = \frac{\log \text{ (number of references with single page locations)}}{\log \text{ (number of page locations of most popular index entry)}}$$

clearly, the greater the number of single page location index entries and the fewer the number of multiple page location index entries, the greater the estimate of s and hence the greater the amount of information per effort under Mandelbrot's definition. In the extreme case, where each index entry refers to one, and only one, page location, Mandelbrot information is infinite. Although we have found so such indexes in the Fondren sample, it is well to note that dictionaries take this form: each main entry occurs once and the referent information is conveniently packaged with the main entry itself rather than through a page location to some other source.

The values of s for each index in the subsample are listed in Table 4.4 in decreasing order of s. Earlier in this chapter we organized the various monographs by LC class and then by total number of entries per monograph. Under this measure the LC classes fell into three disjoint sets corresponding roughly to the descriptive materials, the technique materials, and the arts. The average slope for each of these three groups are respectively, 2.83, 2.37, and 2.79. The differences between the means are not only insignificant statistically; they do not even provide a corresponding ordering, were they significant. Thus the slope (and hence Mandelbrot's measure of average information per average effort) provides an independent measure of the index.

The 28 values of s are plotted in Figure 4.3 on log-normal paper. The distribution of values is reasonably approximated by a straight line as might be expected since as we have now shown, s is a normalized measure of information.

However, except for specialized indexes such as dictionaries, multiply occuring entries do occur, thus depressing the information ratio. For the sample plotted in Figure 4.3, the average value of s is 2.66, quite close to the natural constant, e, which is 2.718. As these multiply occurring entries do reduce the information ratio by increasing the effort required, it is appropriate to inquire as to what role they play in the index.

Some hint as to the nature of this phenomenon can be obtained by examining the role of the multiply occurring entries in the context that Zipf first studied them;

Table 4.4

ZIPF-MANDELBROT EXPONENT
FOR INDEX LOCATION DISTRIBUTION

LC Number	S
PT7244	5.47
QD9	4.43
HB199	4.33
BV2532	3.81
DA690	3.54
TK153	3.53
E741	3.39
Q391	3.20
QA303	3.06
E178	2.81
QL703	2.71
RM721	2.71
BF181	2.69
DF521	2.64
Z5782	2.49
ND553	2.26
HM66	2.15
F864	2.08
PR2831	1.96
LB875	1.94
LC191	1.93
HF2046	1.86
D443	1.81
HD20	1.71
PR5588	1.69
PN2598	1.67
DS423	1.43
JA84	1.09

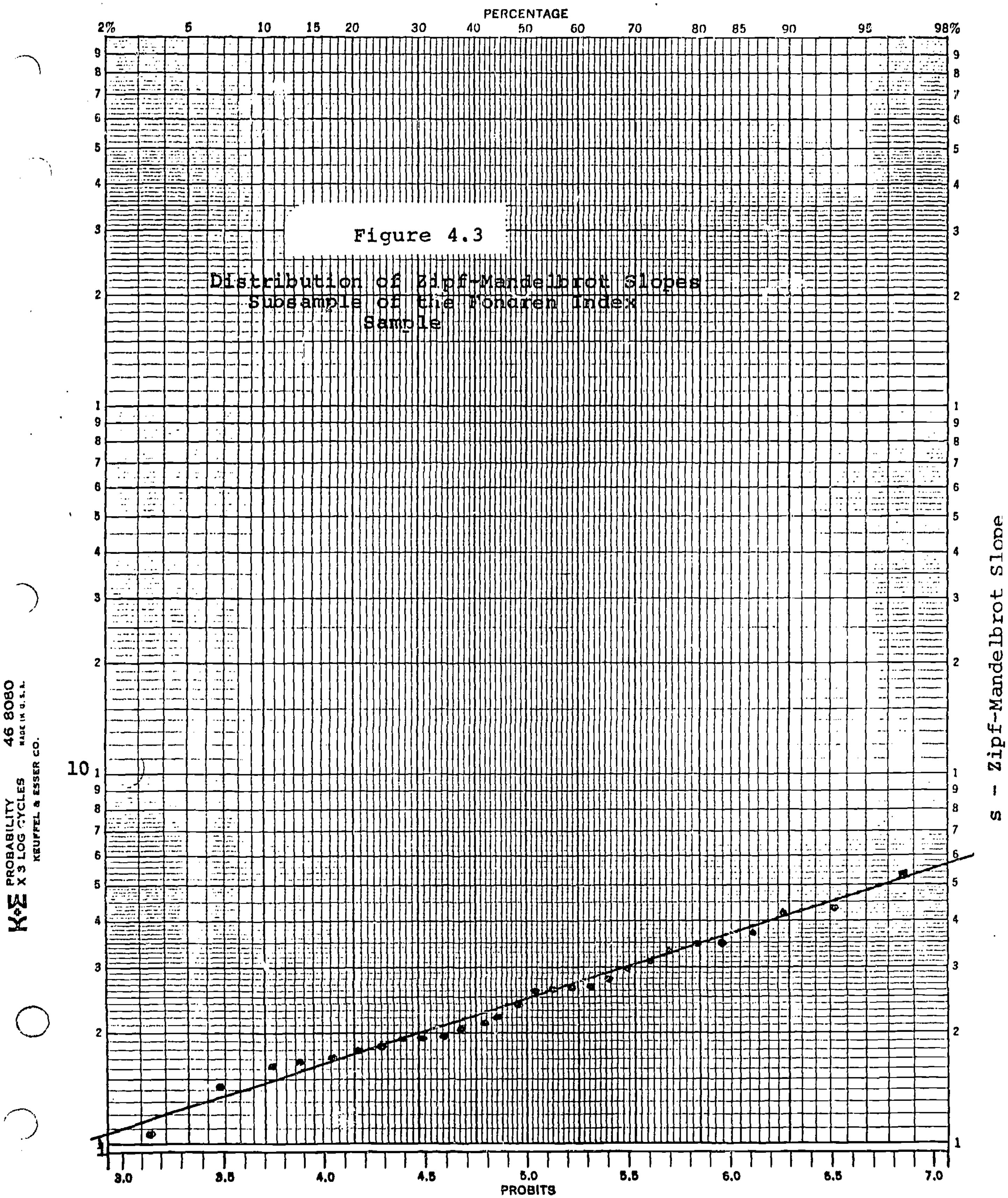


Figure 4.3

Distribution of Zipf-Mandelbrot Slopes
Subsample of the Fondren Index
Sample

in natural language itself. Even a cursory examination of a frequency ordered word list such as those prepared by Thorndike and Lorge (4) and Kucera, et al, (5) is sufficient to show that the most frequently occurring entries are the structure words (i.e. words with parts of speech other than noun, verb, adjective, and adverb). Such words provide the structure in which the information is embedded, but do not, at least in the broad sense, contain information themselves. Except for the rare case (e.g. in the use of certain prepositions in mathematical treatises) such words almost never occur in first position in an index entry.

In this context, it seems natural to suggest that the index entries that occur with many page locations play a fundamentally different role from those that refer only to one or a few page locations. Roughly speaking, we might say that the multiply occurring entries carry the semantic structure in much the same way that the multiply occurring words carry the syntactic structure. Suppose, for instance, that the term California appears in an index with, say, 15 page locations. It would seem reasonable to conclude, even with no other information about the accompanying text, that the text is very much concerned with California in a global manner. Reference to each of the various page locations would presumably uncover a variety of bits of information about California and in this particular sense, we could say that California was one of the "subjects" discussed in the book. If on the other hand, we were to find another book, say on population statistics, whose index contained a single page location for California, it would seem appropriate to conclude that California was one of many items discussed in the text rather than a main subject of the text.

In short, if one is interested in "population statistics for the state of California" one can either go to a book on population statistics and look in the index for California, or one can go to a book on California and look in the index for population statistics. For obvious reasons both types of information packaging exist and access to the packaged information is generally, though not always, provided both ways: by subject to allow the user to get to the proper book, and by index entry to allow the user to obtain the specific fact once he has gotten to the proper book.

The multiply occurring entries thus provide a sort of transition from the "specific fact" aspect of the problem to the "general subject" aspect of the problem. They provide the basis for an algorithmic identification of the semantic structure in the same way that the structure words provide a basis for the algorithmic identification of the author's syntactic style. (See Mostellor and Wallace (6))

For both the word frequency distribution and the index page location distributions, there is no clear break between the set of frequently occurring items and the set of non-frequently occurring items. However, the previously developed arguments on the access level structure provide a technique for establishing break points in the distribution: the set of most frequently occurring entries can be defined as 1/900th of the whole set of entries. This has been done for the subsample of indexes from the Fondren sample. The results are tabulated together with the LC class, the LC subject headings, and the title in Table 4.5.

Looking first at the subject heading and title information in Table 4.5, it is clear that approximately two-thirds of the subject headings are direct transformations (through the subject heading authority list) of the title information. This observation, of course, sheds considerable insight into the discussion of the utility of permuted title indexes: anything as cheap as a permuted title listing that can supply in the order of two-thirds of the subject heading information automatically is clearly useful. At the same time a device that misses one-third of the potential information is clearly not sufficient.

In this context the role of the multiply occurring index entries becomes more obvious: most of LC subject headings that are not derivable from the title information are derivable from the multiply occurring index entries either directly (e.g. Andalusite, U.S.A. vs. Andalusite) or at a higher level of synthesis (e.g. gaseous discharge tube + ultra violet light + reaction, reactors vs. electrical apparatus and appliances). At this stage it is not necessary to re-open the much discussed question of whether classification of documents can be obtained economically through purely algorithmic processes; other simpler problems must be solved first (e.g. the automatic derivation of the index itself). However, it is essential to obtain a clearer understanding of how the various access devices already in operation interact with one another. The preliminary results derived from Table 4.5 make it clear that there is a direct relation between the LC subject headings, the monograph titles, and the multiply occurring index entries. The utility of title derived indexes is manifest by their present use and persistence. It remains to determine the utility of

Table 4.5

COMPARISON OF HIGH-FREQUENCY INDEX ENTRIES
WITH LC SUBJECT HEADINGS & TITLES

LC Class	1/900th Entires	LC Subject Headings	Title
BF181	1. Marston, W. M. (Author) 2. Freud, Sigmund	1. Psychology, Physiological	Integrative Psychology
BV2532	1. Fallen, The	None	The History of the Foreign Missionary Society
D443	1. Great Britain mentioned	1. Europe-Politics-1914	Ten Years of War & Peace
DA690	1. Sackville, Lady Margaret (afterwards Countess of Thanet) mentioned in Lady Anne Clifford's Diary	1. Knole Park, Sevenoaks, Engl. 2. Sackville Family	Knole and the Sackvilles
DF521	1. Churches: in Constantinople 2. Frescoes	1. Byzantine Empire-Civilization	Byzantium
DS423	1. Krsna 2. Siva 3. "Bhagavad-Gita" 4. Visnu 5. Brahman 6. Guru(s)	1. India-Civilization	The Cultural Heritage of India
E178	1. Beard, Charles A. & Mary 2. Jefferson, Thomas 3. Turner, Frederick Jackson	1. U.S.-Hist.-Addresses, Essays, lectures 2. U.S.-Hist.-Historiography	Understanding the American Past

Table 4.5 (Continued)

LC Class	1/900th Entries	LC Subject Headings	Title
E741	1. Prices: agricultural 2. Foreign Relations: Anglo-American 3. Federal Income Tax: individual 4. Tax: individual income 5. Farmers, income of 6. Legislation: agricultural 7. Agricultural, legislation for 8. Railroads: rates of	1. U.S.-Hist.-20th cent.	American Epoch
F864	1. Mass	1. Ansa (sic!) Juan Bautista de 2. California-descr. and travel 3. San Francisco-Hist.	Anza's California Expedition
HB199	1. Terborgh, Gene 2. Breakeven Charts, Examples of	1. Economics	Engineering Economy
HD20	1. Charts on simulated business results	1. Operations Research 2. Industrial Management-Research	Operations Research for Industrial Management
HF2046	1. Chamberlain, J.	1. Free trade and protection --Free Trade. 2. Tariff--Gt. Brit.	The Return to Protection
HM66	1. Trade Unions	1. Sociology	Social Theory
JA84	1. Economy	1. Political Science-Hist.-Russia	Russian Political Thought

Table 4.5 (Continued)

LC Class	1/900th Entries	LC Subject Headings	Title
LB875	1. America	1. Education 2. Literature-Study and Teaching	Two Views of Education
LC191	1. Children, Disease of	none	Education and Social Progress
ND553	1. "Bride Stripped Bare By Her Bachelors, Even, The" (Ducamp)	1. Duchamp, Marcel, 1887- 2. Cage, John 3. Rauschenberg, Robert, 1925- 4. Tinguely, Jean, 1925-	The Bride and the Bachelors
PN2598	1. Butler, Pierce	None	Fanny Kemble
PR2831	1. Greg, Walter	1. Shakespeare, William. Romeo and Juliet 2. Shakespeare, William-Bibl.-Quartos	The Bad Quarto of Romeo and Juliet
PR5588	1. Keats, John	None	Theme and Symbol in Tennyson's Poems to 1850
PT7244	1. Bjark: Bjarkamal, anon.	1. Scalds and Scaldic Poetry 2. Icelandic and Old Norse Poetry	Den Norsk-Islandska Skaldediktningen
QA303	1. Cauchy, A. 2. Euler	1. Calculus	Vorlesungen Uber Differential und Integralrechnung

Table 4.5 (Continued)

LC Class	1/900th Entries	LC Subject Headings	Title
QD9	1. Gregory	1. Chemistry-Bibl. 2. Reference Books	Library Guide for the Chemist
QE391	1. Researves, India 2. Andalusite, U.S.A.	1. Sillimanite 2. Andalusite 3. Cyanite	Sillimanite
QL703	1. Carnivore 2. Bat(s)	1. Mammals	Principles in Mammalogy
RM721	1. Muscle Contraction	1. Gymnastics, Medical	Therapeutic Exercise
TK153	1. Tube, gaseous-discharge 2. Hertz 3. Light, ultra-violet 4. Maxwell 5. Reaction, Reactors 6. Valence electrons	1. Electrical apparatus and appliances 2. Electrons	Electrons at Work
Z5782	1. Passion 2. Comedy 3. Latin 4. Staging	1. Drama, Medieval-Bibl.	Bibliography of Medieval Drama

index entries, over and above their obvious utility in providing access to a book's contents, once the book is in hand. This question will underlie much of the discussion in the next two chapters.

Before turning to this question, however, it is useful to shed some light on how the indexer controls the multiplicities in the index and hence the value of s and the shape of the particular entries that will receive the highest numbers of page locations. Obviously, this can be done in several ways involving such delicate questions as the determination of how the indexer decides whether a particular word, or sequence of words, on a particular page should rate an entry in the index. At a simpler level, the indexer has the opportunity to reduce multiplicities by increasing the length of the entry. Thus in a work on history, the indexer can either provide a single entry for *war*, with a large number of multiplicities, or he can break this same set of entries down into subsets involving particular wars such as *civil war*, *world war*, etc.

That this mechanism is in fact used is easy to demonstrate. Table 4.6 provides the frequency distribution for the 27,188 index entries in a uniform random subsample of 35 indexes in the Fondren sample by word length. As might be expected, the distribution can be reasonably approximated by a lognormal distribution as shown in Figure 4.4. The arithmetic mean of this distribution is 3.68 words per index entry. Only 13.5% of the entries are one-word entries. This is somewhat larger than the 9.1% found in a smaller sample of indexes to statistical books studied by Dolby (7) but still provides strong support for the hypothesis advanced in (7) that the great bulk of the entries in back-of-the-book indexes are multi-word entries.

This observation has considerable significance for the design of automatic indexing procedures. If one-word entries constitute only 13.5% of the total index, it seems unlikely that detailed frequency studies of words will provide much insight into the problem of deriving index entries automatically. In some of the earliest work on this subject, Luhn (8) attempted to derive indexes from word frequency counts, with limited success. More recently, Damerau (9) established a procedure for deriving coordinate index terms (to be used later via machine searches) based on word frequency counts. Bloomfield's (2) study of Damerau's procedure makes it clear that coordination of the single terms derived by Damerau rarely leads to an index entry derived by humans for the same material. As we shall show in the next chapter, there is more to be gained by deliberately suppressing the one-word entries, rather than by attempting to emphasize them.

Table 4.6

Distribution of Index Entries by
Word Length - Subsample of
The Fondren Index Sample

Number of Words	Number of Entries	Cumulative Number	Cumulative Percentage
1	3673	3673	13.51
2	6563	10236	37.65
3	4817	15053	55.37
4	3905	18958	69.73
5	2839	21797	80.17
6	1969	23766	87.41
7	1243	25009	91.99
8	801	25810	94.93
9	516	26326	96.83
10	281	26607	97.86
>10	581	27188	100.00

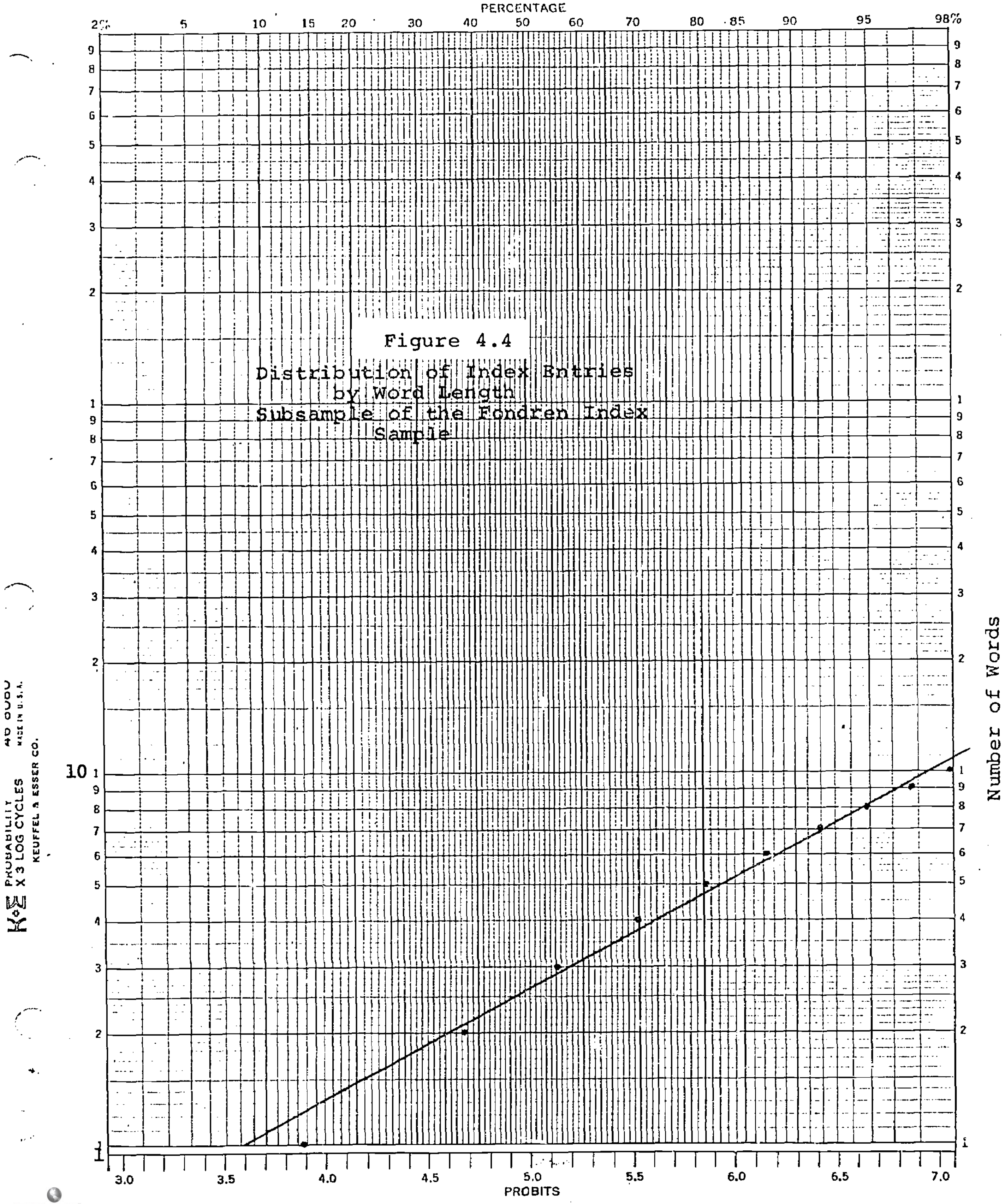


Figure 4.4

Distribution of Index Entries by Word Length Subsample of the Fondren Index Sample

The observation that index entries are usually one-word entries also has some impact on a variety of questions involved with the use of indexes in agglomerated form. This will be discussed at some length in Chapter VI.

References

1. Dolby, J. L., H. L. Resnikoff, and V. Forsyth, *Computerized Library Catalogs: Their Growth, Cost, and Utility*, M.I.T. Press, Cambridge, 1969.

2. Bloomfield, Masse, "Evaluation of Indexing, 3. A Review of Comparative Studies of Index Sets to to Identical Citations", *Special Libraries*, December 1970, 554-61.

3. Kendall, M. G., "The Bibliography of Operational Research", Operational Research Quarterly, 2(1960), 31-6.

4. Thorndike, E. L., and I. Lorge, *The Teacher's Word Book of 30,000 Words*, Columbia University, New York, 1944.

5. Kucera, H., and W. N. Francis, *Computational Analysis of Present-Day American English*, Brown University Press, Providence, Rhode Island, 1967.

6. Mosteller, F., and D. Wallace, "Inference in an Authorship Problem," *Journal of the American Statistical Association*, 58(1963), 275-309.

7. Dolby, J. L., "The Structure of Indexing: The Distribution of Structure-Word-Free Back-of-the-Book Entries", *Proceedings of the Annual Meeting of the American Society for Information Science*, 5(1968), 65-72.

8. Luhn, H. P., "A Statistical Approach to Mechanized Encoding and Searching of Literary Information", *IBM Journal of Research and Development*, 1(1957), 309-17.

9. Damerau, F. J., "An Experiment in Automatic Indexing", *American Documentation*, 16(1965), 283-9.

CHAPTER V

ALGORITHMIC TEXT INDEXING

ALGORITHMIC TEXT INDEXING

An index increases access to a particular corpus of information. Until recent times most indexes followed the text material in certain types of books. Although this may still be true today, the emphasis of research into the nature of indexing has shifted to indexes of other types of corpora, such as the permuted title index and its variants and the citation index, which index collections of document titles rather than the text of the documents. Indeed current information retrieval efforts appear to exclude consideration of back-of-the-book indexes. For instance, Salton (1) offers a brief discussion of term-oriented, or derived indexes, of which the back-of-the-book indexes are usually instances, but the applications he describes are to collections of document titles. *The Encyclopedia of Linguistics, Information and Control* (2) mentions only citation indexing.

This chapter is also exclusively concerned with back-of-the book indexes; hereafter the term *index* will be used in this restricted way.

The principal result presented here is an algorithm for the automatic construction of an index from run g text in machine readable form. A preliminary versi of the algorithm was implemented by hand and used t derive the index to Dolby, Forsyth, and Resnikoff (3). The version presented here has been programmed for the IBM 360/30 using a set of assembly code macros and tested on a set of 50 abstracts of statistical papers published in the *Annals of Mathematical Statistics* and a second set of abstracts published in *Cancer Research*.

The difficult question of determining what is to constitute an adequate index for a given corpus of running text is not considered here, although reference is made to an earlier study (Dolby (4)) that considered certain obvious statistical characteristics of published indexes as well as to the previous chapter.

The cost of deriving the index entries and formatting them into standard format is approximately 2¢ per line of input text, based on standard commerical rates (west coast of the United States).

Let us assume that an index is an ordered collection of word sequences (or transformations thereof) from the running text together with appropriate locator designations (e.g. page numbers). A reasonable first step in deriving such an index is to partition the text into a set of word sequences using, in this case, marks of punctuation and structure words to determine the sequence boundaries.

Each sequence is then examimed to determine whether it should be deleted from the set. In particular, sequences consisting of structure words only are deleted. For reasons that will become evident later, sequences consisting of single words and sequences that occur only once in the entire corpus are also deleted from the set.

Of the various possible transformations it is obviously desirable to identify singular and plural forms, to invert certain word sequences (at least selectively) so as to provide access to words occurring only at the end of the word sequences, and to superimpose a "see" and "see-also" facility to permit more complex transformation.

Implementation of such an algorithm requires repeated access to various lists of words and morphemes. Computer time will obviously be strongly influenced by the strategies employed to accomplish these comparisons. To cite the most obvious example, it is clearly more efficient to store the list of structure words (which is relatively small but contains many words of high occurrence frequency) rather than the list of content words which has the converse properties.

Where possible, significant gains can be made by testing for word classes rather than for individual words. Thus, it is useful to identify all participial forms as these do not generally appear as index entries. On the other hand, provision must be made to allow the override of such rules for cases of particular importance. (e.g. stratified sampling is an important statistical entry that should not be suppressed.)

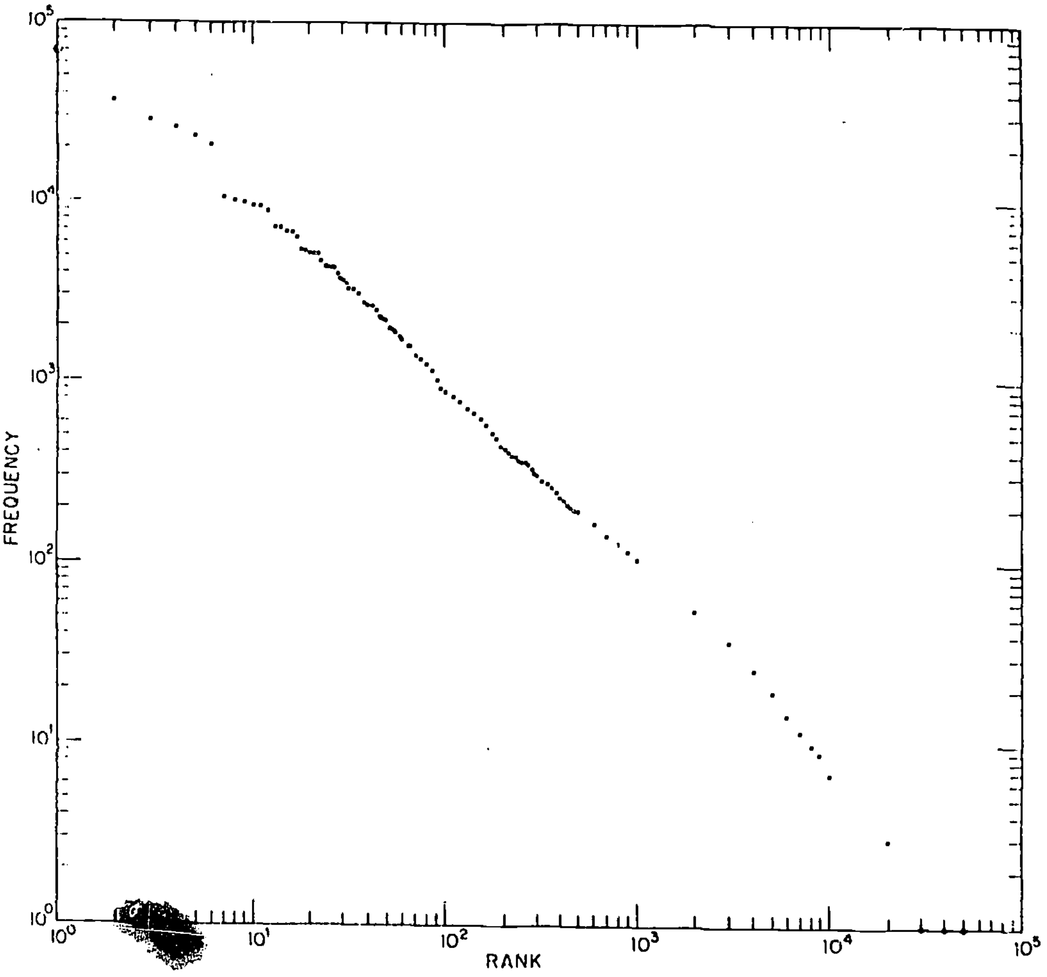
As the function of these various lists is primarily to delete words from the index, it is convenient to refer to the lists as "stop" lists and the sets of override words as "go" lists. Although sufficient testing on a wide variety of subject matter is not yet available, it would appear that the stop lists are basically independent of subject material and the go lists are subject

Figure 5.1

Word Frequency versus Rank

Brown University Standara Corpus of

American English



dependent. Thus a careful study of available authority lists in the subject field would be necessary to insure proper operation of the algorithm. (Such a study would be necessary in any event to prepare the "see" and "see-also" entries.)

Preliminary segment boundaries are established by marks of punctuation (other than the hyphen and apostrophe). Within the segments thus established, further boundaries are introduced between sequences of consecutive stop words (see Table 5.1) and non-stop words. As a simple expedient, all words in the stop list ending in *s* have the *s* removed and the match between the current word and the stop list is made after the final *s* (if any) has been removed from the current word. More sophisticated "plural logic" would be justified here only if the stop list were expanded substantially and in its expanded form contained a significantly larger number of "irregular" plurals.

The selection of the words to be used in the stop list provides an intriguing problem. Clearly, all structure words (neglecting archaic forms) should be included. It turns out also to be useful to include high frequency adjectives and verbs. It is therefore tempting to simply select the first n words from a rank ordered word frequency list. Unfortunately, there is no clear break in such a list in the vicinity of a reasonable cutoff (see Figure 5.1). Thus the cutoff must be made simply in terms of finding a reasonable trade off between added machine costs in testing against large lists, and added editing costs at the other end due to failure to suppress words. Based on the developments of Chapter II, we would expect the cutoff to be in the order of 1/30 of the vocabulary. The word list used here has been purposely kept short during programming and should probably be expanded by a factor of two or three in actual use.

The list organization as presently implemented is also quite simple: as the word length (in characters) of the current word is known at the time of the match, the list is broken down by word length and arranged alphabetically, within the sets of each length. Matching is done sequentially with termination on a match or when the current word is low to the list. Expansion of the lists would probably make it useful to use a hashing technique.

The next segmentation stage consists of segmenting the sequences of non-stop words into consecutive sequences of words ending in *ed*, *ly*, *ing*, or *ful* and sequences of

TABLE 5.1

SHORT LIST OF STOP WORDS
ARRANGED BY WORD LENGTH

an
at
be
by
do
go
ha
he
hi
if
it
on
or
me
my
no
so
to
up
wa
we

all
and
any
are
but
can
did
due
few
for
get
had
her
him
how
let
may
new
not
now
off
old
one
our
out

own
put
see
she
the
thi
too
two
way
who
you

also
back
been
both
come
down
each
even
from
give
good
have
here
hold
into
just
know
last
lead
lend
like
long
made
make
many
more
most
much
must
only
over
part
said
same
show

some
such
take
tend
term
than
that
them
then
upon
very
well
were
what
when
will
with
work
your

about
above
admit
after
among
begin
could
drawn
first
found
given
great
imply
known
might
never
other
refer
right
sense
shown
since
still
their
there
these
those

three
under
until
usual
where
which
while
whose
wider
would
yield

become
before
behave
better
cannot
change
chosen
denote
depend
derive
discus
either
extend
higher
implie
little
permit
relate
reduce
result
second
should
unique
variou
wherea
within

against
another
because
between
certain
consist
earlier
further

general
improve
include
instead
operate
present
previou
provide
require
several
similar
special
through
unknown
without

consider
original
possible
satisfie
together

arbitrary
different
excellent
important
otherwise

additional
elementary
particular

N.B. All one-letter words are stopped. Terminal s is removed, thus ha stops has .

words not ending in any of those four suffixes. The current go list to override this segmentation consists of only three words (*family*, *stratified*, and *sampling*) and is included only to insure that the facility exists in the program.

The structure words *of* and *in* are not included in the main stop list so as to allow sequences such as *analysis of variance* and *convergence in measure* to emerge as index sequences. However, it is clear that primary index entries do not include entries beginning or ending with *of* or *in*. Hence the final segmentation step is to segment beginning or ending occurrences of these to words from the non-stop, non-(*ed*, *ly*, *ing*, *ful*) word sequences.

Following a suggestion of John Tukey, we have investigated the utility of "stopping" all short words, i.e., words with fewer than n characters. Such a procedure would clearly speed up the program and set aside the difficulty of running down a number of short words that occur with sufficient frequency so as to be included in a reasonable system, (such as those occurring in Latin phrases). Based on present experience, it appears that suppressing words with fewer than four characters is reasonable. This procedure has been used in the experimental run on the 50 abstracts from *Cancer Research*, but not on the two earlier examples presented here.

All segments other than those consisting wholly of non-stop, non-(*ed*, *ly*, *ing*, *ful*), with beginning and ending *of* and *in* removed, are deleted. Of the segments remaining, all segments consisting of single words are also deleted. Experimentation with this step in the procedure stems from an observation made in Dolby (4) that one word entries in published indexes occur with surprisingly low frequency. Hence, the obvious strategy is to suppress all entries with exceptions rather than to pass all with exceptions.

The override to single-word suppression can take several forms. First, a go list can be appended (though none is used in the present implementation). *Independence* would be an obvious choice for statistical subject matter. Second, proper names, that is, words in all caps or initial caps could be used as an override. (This was done in the manually implemented version used on Ref. (3) but has not been exercised in the machine implementation.) Finally, single-word primary entries can, and do occur in the inverted entries studied below.

This reduced list of segments, or possible index entries, must now be transformed in certain obvious ways both to achieve proper compression in the final index and to provide at least the appearance of a manually prepared index. One obvious consideration involves the problem of identifying singular and plural forms. Again, a relatively simple strategy is sufficient to take care of most of the problem. Plural forms are rarely used as modifiers and when so used are used with a high degree of consistency. Thus if *least squares method* occurs, it is highly unlikely that *least square method* will also occur (though *least squares methods* might well occur). Hence it is only necessary to prepare for plurals that occur at the end of the entry.

The most frequently occurring plural form is obtained by adding *s* to the singular form. If the final *s* is replaced by a code that will sort immediately after blank (but prior to *a*) it is possible to compare successive entries after sorting and to eliminate the final *s* from all entries that follow entries that are otherwise identical. The final *s* is then restored in all other cases. In the application to the statistical abstracts 311 of the 946 entries ended in *s*. Of these, 41 were stripped of the final *s* to provide the required identification. More sophisticated logic of the same variety could be added to handle plural forms such as *processes*, *densities*, and *matrices* although a quick survey of the 946 entries disclosed only four such occurrences where identification was desirable.

Another purely manipulative step that must be introduced at this stage is the generation of inverted entries to provide access to words occurring at the end of the text ordered entries. There appear to be two main forms of interest. The first, typified by *analysis of variance*, can be implemented by the obvious algorithm that produces *variance, analysis of*. A more sophisticated form could be used to suppress one or the other of the two variants. A pair of relatively short, subject dependent, stop lists would probably suffice for this purpose.

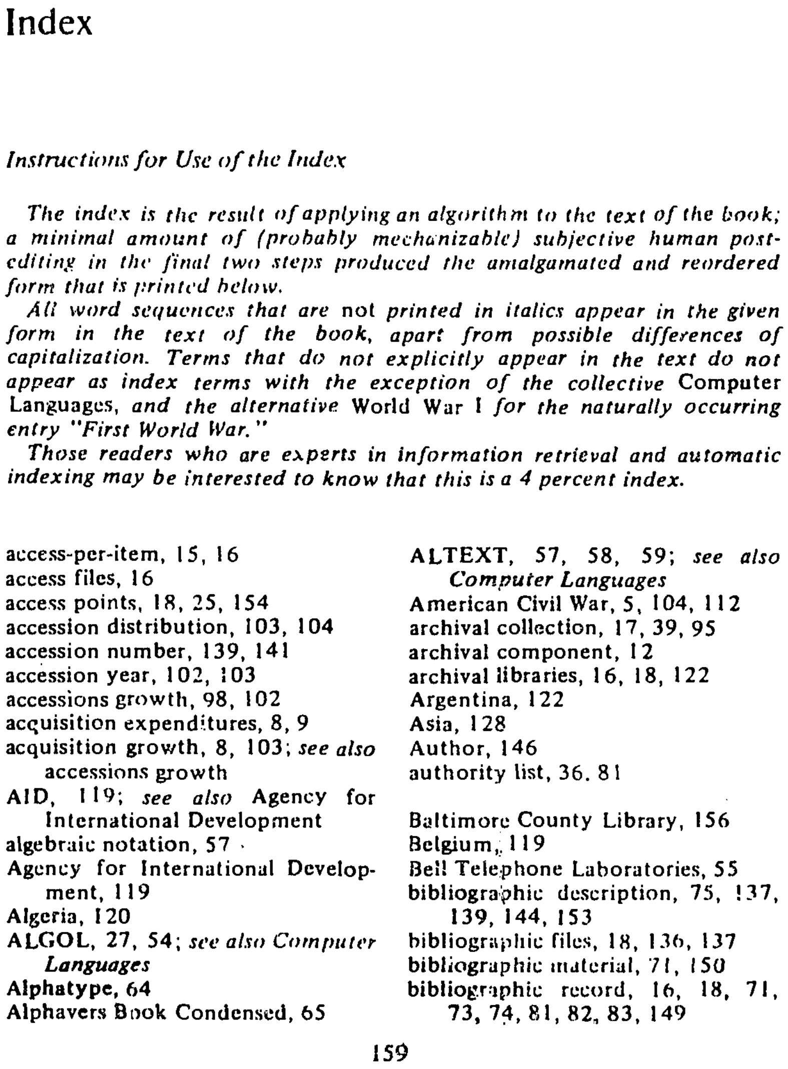
A second type of inversion, typified by mapping *normal distribution* into *distribution, normal* could either be implemented by a go list of modest proportions or by ordering the entire set of entries by last word and then inverting all sets involving a common last word of sufficiently high frequency. Neither of these alternatives have been tried at this time, though some statistics have been gathered on the behavior of statistical terms from this point of view.

In addition to the deletion of one-word entries, it is evident, when one operates on full text, that it is entirely safe, and indeed quite useful, to delete entries that occur only once in the text. Intuitively, one can argue that if a term is not mentioned at least twice (allowing for plural variants and the like) then there is little likelihood that enough information is presented about that entry to make it worthwhile as an entry in the final index. Practically, an examination of singly occurring entries in the samples we have studied thus far makes it clear that this is a highly useful device for eliminating much of the "noise" that inevitably is present when one takes such a simple view of English syntax. Statistically, the step can be justified on the grounds that the resultant index is of the proper size (as a percent of the volume of the book indexed) when such entries are left out, but noticeably too large if they are left in.

The use of this device must be tempered by knowledge of the text. For instance, this device was not used in the index to the statistical abstracts, as it was evident that the abstracts did not possess sufficient redundancy to allow proper operation of such a mechanism within an abstract, and it seemed unwise to base the use of such a mechanism on a (not necessarily homogeneous) set of abstracts. Presumably there are certain books whose text has a very low redundancy; for these this type of deletum should not be impletmented.

The manual implementation of the algorithm on book length material (reference (3)) is shown in Figure 5.2. Two systematic departures from the general algorithm were made in implementing it: first, names of States were systematically deleted from the index; second, a list of special words for inclusion in the index was used, containing names of countries and languages. Both decisions insure uniformity of in- or ex- clusion of terms in each class without regard to the relative significance of each usage. Finally, as described in the Instructions for Use of the Index, two index terms were manually inserted: the collective Computer Languages, and the alternative World War I for the algorithmically occurring First World War.

Perhaps of greater theoretical interest than those terms that appear in the index in Figure 5.2 are those terms that were deleted by the requirement that each entry that appears in the index, except for entries having special format properties, refer to more than one location in the text. Table 2.4 lists those word sequences which were excluded from the index for this reason. Preceding some of the words are letters which describe properties of the word sequence: 'p' indicates that the sequence is a plural form of another word sequence selected by previous steps of the algorithm; the plural sequence is therefore equivalent to the singular one, and hence appears in the final index. Sequences preceded by 'i' appeared in italic type font. It appears that this font



Index

Instructions for Use of the Index

The index is the result of applying an algorithm to the text of the book; a minimal amount of (probably mechanizable) subjective human post-editing in the final two steps produced the amalgamated and reordered form that is printed below.

All word sequences that are not *printed in italics appear in the given form in the text of the book, apart from possible differences of capitalization. Terms that do not explicitly appear in the text do not appear as index terms with the exception of the collective* Computer Languages, *and the alternative* World War I *for the naturally occurring entry "First World War."*

Those readers who are experts in information retrieval and automatic indexing may be interested to know that this is a 4 percent index.

access-per-item, 15, 16
access files, 16
access points, 18, 25, 154
accession distribution, 103, 104
accession number, 139, 141
accession year, 102, 103
accessions growth, 98, 102
acquisition expenditures, 8, 9
acquisition growth, 8, 103; *see also* accessions growth
AID, 119; *see also* Agency for International Development
algebraic notation, 57
Agency for International Development, 119
Algeria, 120
ALGOL, 27, 54; *see also Computer Languages*
Alphatype, 64
Alphavers Book Condensed, 65
ALTEXT, 57, 58, 59; *see also Computer Languages*
American Civil War, 5, 104, 112
archival collection, 17, 39, 95
archival component, 12
archival libraries, 16, 18, 122
Argentina, 122
Asia, 128
Author, 146
authority list, 36. 81

Baltimore County Library, 156
Belgium, 119
Bell Telephone Laboratories, 55
bibliographic description, 75, !37, 139, 144, 153
bibliographic files, 18, 136, 137
bibliographic material, 71, 150
bibliographic record, 16, 18, 71, 73, 74, 81, 82, 83, 149

159

Figure 5.2

ALGORITHMIC INDEX TO REFERENCE (3)

bibliographic tool, 135, 144
bibliographical reference, 69, 126
bold face, 63, 64, 65, 68
book catalogs, 1, 31, 69, 88
Book Condensed, 65; *see* Alphavers Book Condensed *type font*
books-per-billion ratio, 123, 124
Brazil, 122
British Museum, 65, 115
Bro-Dart catalog, 69
Bulgaria, 119

Canada, 119, 122, 123, 124
capitalization and other errors, 81
card catalog, 1, 25, 73, 87, 93, 114, 135
cards per catalog drawer, 89
catalog
  automation, 15, 37, 39
  card, 41, 44, 89, 92, 93, 95
  conversion, 40
  entry, 39, 68, 72, 75, 149
  file, 137, 138, 143, 153, 157
  information, 23, 36, 40, 155, 157
  material, 70, 154
  mechanization, 16
  operation, 26, 39, 156
  record, 18, 114, 142, 144
  system, 1, 114
  use, 24
characters per record, 17, 43
characters per square inch, 61, 62
Chemical Abstracts, 116, 124
China, 128
Church, 117; *see* Roman Catholic Church
CHY, 41, 43; *see* Columbia-Harvard-Yale Medical Libraries
circulation rate, 12, 130
Civil War, 112, 113; *see* American Civil War
class field, 141, 142, 143, 144, 146, 148, 152
class number, 95, 104, 142
COBOL, 25, 27; *see also Computer Languages*
collection growth, 98, 102
collection size, 89, 137, 143
Columbia-Harvard-Yale Medical Libraries, 41; *see* CHY
columns per page, 68, 69
COMIT, 53, 54, 55, 56, 57, 58, 59; *see also Computer Languages*
COMIT II, 53, 54; *see also Computer Languages*
comparative input equipment costs, 45
composite figures, 43, 44
composition cost, 69, 153
Computer Languages
  ALGOL, 27, 54
  ALTEXT, 57, 58, 59
  COBOL, 25, 27
  COMIT, 53, 54, 55, 56, 57, 58, 59
  COMIT II, 53, 54
  FORTRAN, 25, 27, 54
  LISP, 56
  LISP 1.5, 56
  PL-1, 27, 71, 72
  SNOBOL, 27, 28, 55, 56, 57, 58, 59
  SNOBOL 3, 55, 56
  SNOBOL 4, 55, 56
  TEMAC, 28, 56, 57
  XPOP, 57, 58, 59
conversion costs, 16, 17, 40
conversion project, 16, 18
corporate authorship, 93
cost
  comparative input equipment, 45
  composition, 69, 153
  conversion, 16, 17, 40
  machine, 29, 31, 40
  operation, 31, 157
  personnel, 15, 16, 31
  quality, 43
  software, 25, 38
cost per character, 16, 44
cost reductions, 36, 51
Costa Rica, 123
county library system, 152, 156
Cyrillic, 75
Czechoslovakia, 119

data base, 78, 126, 127
data points, 9, 89

East Germany, 119
economic growth, 8, 15, 115, 128
economic indicators, 19, 115
economic statistics, 117
edit lists, 43, 44
electronic composition, 17, 19, 31, 70
Embassy, 119
English, 94, 127, 128, 131, 143
English language share, 128
error-detection, 43, 71, 73, 74
error messages, 71, 74
error rate, 74
Europe, 119
European GNP, 119
exponential growth, 8, 15, 16, 29, 95, 100, 104, 130, 144, 153

Finland, 117
First World War, 112
Fondren, 89, 92, 93, 104, 112, 131, 132, 133, 152; *see* Rice University
Fondren Library, 5, 87, 89, 100, 102, 122
Fondren Sample, 89, 100, 104, 131, 150; *see* Sample
Fondren shelf list, 89, 92, 95
FORTRAN, 25, 27, 54; *see also Computer Languages*
France, 120
French, 119, 120, 124, 131
French GNP, 120

German, 131
Germany, 127, 148
GNP, 98, 102, 103, 116, 117, 119, 120, 121, 122, 123, 124, 126, 127, 128, 130, 132; *see* gross national product
graphic arts quality, 32, 64, 153, 154
Great Depression, 5, 113
Great War, 2, 5; *see* First World War
gross national product, 8, 12, 15, 98, 100, 103, 116, 117, 119, 130, 131, 133; *see* GNP
growth rate, 2, 5, 51, 88, 95, 98, 100, 102, 103, 104, 115, 119, 124, 128

Harvard, 71, 75, 78, 82; *see* Harvard University, Harvard sample, Widener Library
Harvard sample, 77, 78, 79, 80, 81
Harvard University, 2, 75
Hebrew, 121
History of the United States, 104
Honduras, 123

IBM
  Selectric Typewriter, 64
  360, 57
  709-7090-7094-7040-7044, 53
  1401, 41, 75
  7094, 41, 57
imprint date, 2, 98, 102, 104, 116, 131
imprint date distribution, 5, 103
imprint distribution, 5, 104
imprint growth, 98, 103
India, 128
Indonesia, 128
information density, 61, 68, 69
information explosion, 5, 12
information retrieval, 33, 93, 130, 140
Italian, 119, 120, 131
Italy, 120
item field, 141, 142, 143, 144, 146, 152

Japan, 127, 128
Japanese, 127, 128
journal title, 126
justification losses, 68, 69

keyboard operator, 73, 75, 81, 83

LACP, 40, 41, 43, 45; *see* LACPL, Los Angeles County Public Library

LACPL, 51
language distribution, 93, 94, 124, 130
language group, 119, 123, 128, 131, 133
language group GNP, 126, 127, 128
Latin, 117
LC, 92, 93, 104, 112, 113, 116, 117, 120, 121, 122, 126, 128, 131, 132, 133, 144; *see* Library of Congress
LC acquisition shares, 119, 126, 127
LC letter class, 104
legitimate codes, 73
library
  activity, 24, 29
  automation, 23, 87
  book catalogs, 61, 68
  budgets, 16, 51
  catalog, 1, 17, 19, 61, 62, 65, 68, 72, 87, 93, 114, 130, 147, 155
  collection, 25, 88, 95, 103, 104, 130
  growth, 2, 15, 100, 130
  growth rates, 1, 12, 31, 124
  management, 116, 135, 136
  operation, 39, 130
  size, 1, 12
  staff, 40, 137, 153
  system, 16, 156
Library of Congress, 2, 17, 92, 94, 115, 120, 121, 122, 124, 131, 136, 143, 148; *see* LC
Library of Congress acquisitions, 116, 119, 131
LISP, 56; *see also Computer Languages*
LISP 1.5, 56; *see also Computer Languages*
Literature, 150
location of holdings field, 152
Lockheed Missiles & Space Company, 57
Los Angeles County Public Library, 40, 51; *see* LACP, LACPL
Luxembourg, 120

machine
  costs, 29, 31, 40
machine (continued)
  language, 54, 55; *see also Computer Languages*
  time, 26, 28, 29, 54, 57, 153
machine-readable
  catalog, 18, 156
  data, 31, 155
  form, 1, 2, 16, 17, 18, 39, 73, 75, 81, 92, 95, 114, 135, 138
magnetic tape, 33, 41, 46
manual methods, 16
MARC, 44, 149; *see* Project MARC
MARC II, 144
mass storage, 33
mathematical journals, 126, 127
Mathematical Reviews, 126
Mexico, 122
Meyer Undergraduate Catalog, 75; *see* Stanford Undergraduate Catalog; *see also* Stanford University
MIT (=*Massachusetts Institute of Technology*), 53
MSU (=*Michigan State University*), 45

national shares, 116
National Union List, 155
Newcastle, 138, 139, 140, 141, 143, *see* University of Newcastle-upon-Tyne
Newcastle Library, 140, 143
nonserial cards, 93, 95
nonserial items, 93
numeric values, 53, 73

OCR (=*optical character recognition*), 39, 45, 46
on-line access, 157
Ontario New Universities Library Project, 40; *see* ONULP
ONULP, 40, 41, 43; *see* Ontario New Universities Library Project
operation costs, 31, 157
order file, 137, 138, 143, 152

order of magnitude improvement, 24, 27
Outline of the Library of Congress Classification, 148

paper tape, 46
personal income, 12
personal name, 93
personnel costs, 15, 16, 31
personnel requirements, 95, 102
Peru, 123
photocomposition, 155, 156
Physics Abstracts, 116, 124
PL-1, 27, 71, 72; *see also Computer Languages*
Poland, 119
Polish, 119
Polish prefix notation, 56
Princeton University implementation *(of SNOBOL)*, 55
Print 68, 31
Project MARC, 153; *see* MARC
public catalog, 88
public library, 8, 9, 12, 16, 24, 39, 51, 137
punch cards, 41, 46
punched card equipment, 44, 46
pure science, 115

quality control, 18, 73, 74, 83
quality costs, 43

random sample, 5, 40, 87, 88, 89; *see also* Sample
Recommending Officer, 121
relative minimum, 5
relative shares, 119
residual errors, 74, 82, 83
retrospective
  catalog, 18, 39, 44
  file, 1, 8, 16, 51, 139, 144
  material, 40
Rice University, 5, 87, 122; *see also* Fondren Sample
Roman, 65
Roman alphabet, 75
Roman Catholic Church, 117
Romania, 119
Russian, 133

Sample, 88, 89, 92, 93, 94, 95, 100, 102, 112; *see* Fondren Sample, Rice University
scientific community, 25, 28
scientific computation, 25, 28
Second World War, 2, 113; *see* World War II
selection field, 139, 140
sequence check, 77
sequence errors, 77, 79
Serbo-Croatian, 117
serial publication, 2, 88, 89
shelf list, 2, 5, 87, 88, 89, 92, 94, 112, 156
size estimates, 2
Slavic, 128
SNOBOL, 27, 28, 55, 56, 57, 58, 59; *see also Computer Languages*
SNOBOL 3, 55, 56; *see also Computer Languages*
SNOBOL 4, 55, 56; *see also Computer Languages*
software costs, 25, 38
sort field, 139, 140
Soviet Union, 119, 127
Stanford, 2, 71, 75, 77, 78, 81, 82, 143, 144; *see* Stanford University; *see also* Meyer Undergraduate Catalog
Stanford sample, 77, 78, 79, 80, 81
Stanford Undergraduate Library, 2, 41
Stanford University, 75, 136
state personal income, 9
statistical data, 5, 116
statistical information, 104, 121, 123
Strike-On Type Faces, 62
subject
  catalog, 2, 147, 154
  class, 142, 148
  field, 141, 146
  information, 148, 149

SUL, 41, 43; *see* Stanford Undergraduate Library, Meyer Undergraduate Library
Switzerland, 119, 120

telephone directory, 68, 69, 155
TEMAC, 28, 56, 57; *see also Computer Languages*
time intervals, 8, 98
title field, 141, 143, 152
title information, 74, 139
title list, 2, 153, 154
titles per subject, 147, 148
trend line, 12
Type Face Design, 61
type face, 61, 62, 64, 65, 68, 70

UC/B, 40, 43; *see* University of California/Berkeley
United Kingdom, 127
United States, 98, 100, 103, 112, 122, 123, 124, 127
United States GNP, 124
United States Gross National Product, 103
university collection, 8, 130
university library, 5, 12, 24, 38, 51, 92, 130, 133, 137, 140, 156
University of California/Berkeley, 40; *see* UC/B
University of Chicago, Graduate Library School, 53
University of Newcastle-upon-Tyne Library, 138; *see* Newcastle
Uruguay, 123
U.N., 119
U.N. Yearbook, 119
U.N. Yearbook of National Account Statistics, 119
U.S. Department of State, 119

Versatile Bold, 65
volumes per card, 88, 94
Volumes per Title, 94

Wall Street Journal, 31
Wang algorithm, 56
West Germany, 127
Widener Library, 2, 155, 156; *see also* Harvard
Widener Shelf List, 75, 148, 155
World Almanacs, 2
World War I, *see* First World War, Great War
World War II, 5, 100; *see* Second World War
World Wars, 5, 124

XPOP, 57, 58, 59; *see also Computer Languages*

TABLE 5.2

Excluded Index Terms Referring to One Location

abnormal parenthesization
absolute frequencies
academic staff
access capability
access point
access system
accessible estimates
p accession distributions
accessions growth rate
acquisition rates
acquisitions data
acquisition mechanisms
acquisition process
acquisition rate
acquisition schedule
acquisition shares
acquisition structure
acquisitions - GNP relation
acquisitions - GNP share equality
acquisitions budget
p acquisitions growth
p acquisitions expenditures
adequate user access
algebraic equations
algebraic expressions
alpha-numeric code
alphabetic code
approximate linearity
approximate normality
p archival collections
archival holdings
archival libraries
archival records
Assembly code programming
"assembly languages"
assignment procedure
author access
author field
author list
author name
author/title list
p authority lists
automated catalog
Auxiliary memory
average cost
average growth
average number
average record length
average time

p Baltimore County Libraries
bedroom states
biases inherent
p bibliographic descriptions
bibliographic holdings
bibliographic indications
bibliographic items
bibliographic listings
bibliographic lists
bibliographic notes
bibliographic practice
p bibliographic records
bibliographic references
bibliographical information
bibliographically incomplete
bibliography
Bibliography Field
bibliography section
book-publication depressions
Book Length
bookseller
budget dollar
budgetary requirements

business community
calculus text
call number

(Canadian)census figures
capital letters
capitalization conventions
capitalization errors
capitalization requirements
card catalog collection
card collection
card files
card space convention
card system
cards per entry
cards per title
"careful" study
case alphabet
catalog card conversion errors
p catalog cards
catalog data
p catalog files
catalog interrogations
catalog preparation
catalog productions

p catalog records
catalog trays
cataloging function
character density·per page
character density per square inch
character manipulation
Circulation
circulation file
p circulation rates
civilization's material aspects

class category
p class fields
i class order
"Collected Works"
Collection Breakdown
p collection sizes
collection subset
common machine
component national growth rates
composite costs
composite estimate columns
p composition costs
composition devices
composition practice
computational ease
computational facilities
computational linguistics
i Computer Line Printer Type Faces
computer programs
computerization costs
Condensed
consecutive years
contents
context permissible
conversion expense
conversion problem
conversion procedure
conversion process
conversion task
copy output cards
core memory
correction capabilities
correction costs
correspondence files
cost
cost area
cost breakdowns
cost equations
cost estimates
p cost factors

p cost figures
cost function
cost increments
cost item
cost levels
cost per title
cost point
cost structure
cost picture
cost study
cost variations
costs per thousand dollars
county library automation projects
p County library systems
county school system
county system
data conversion
data files
data object structure
data objects
i date of access field
i date of order field
decimal classification system
density output
detail level
dictionary lookup procedures
document descriptions
document identification procedures
dollar equivalents
dummy entities
dynamic aspects

economic analysis
economic aspects
economic data
economic depression
economic disintegration
economic references
economic size
economic state
economic statistical data
economic strength
economic units
edition statement
educational advantages
electronic photocomposition
electronic typesetting devices
i elementary calculus
English-language sentences
English-speaking
error-checks

error-correction capability
European
executable statements
expansion ratio
"explosive" growth
exponential curve
exponential expansion
exponential function
exponential imprint date distribution
exponential library growth rates
exponential rate
faculty library committee
feedback response
field names
i fields
fields per record
file figures
file maintenance
file records
file structure
file system
financial data
financial community
financial transactions
(first) generation machines
fixed absolute growth
floating point arithmetic
follow-up correspondence
foreign language acquisitions
foreign language documents
foreign titles
Format-Dependent Errors
format capabilities
format compromise
Format control
format elements
format requirements
French-African
French-speaking
functional collection
fund name
fundamental processes affecting libraries
fundamental structure
future funding needs
geographic area
geometric decrease
global category
global check
global war
GNP-acquisitions relation

GNP at Market Prices
graph paper
graphic arts
graphic representation
Gross Domestic Product
p gross national products
Gross Personal Income
ground-level extension
growth challenge
growth periods
growth phenomenon
growth problems
p growth rates
growth statistics
hardware costs
p Harvard samples
"higher level" languages
historical events
historical significance
human costs
human readable document
identification number
illegitimate code
implementation cost
imprint data
p imprint dates
imprint date growth
imprint decade
p imprint distributions
in-depth studies
in-school access
income data
income growth
income ratios
indented lines
information base
information fields
information per inch
information per page
information run-over
input costs
input errors
input format
input methods
input program
inquiries per record
instruction per second
intercolumn space
interentry blank lines
interlibrary loan service

interlibrary loans
interword spaces
interpretive approach
i item
p item fields
item purposes
items per year
journal-to-language assignment process
p journal titles
key economic issue
key information
key library personnel
key words
keyboard conventions
p keyboard operators
keypunch equipment
labor categories
language acquisitions
language count
language expertise
Language Field
p language groups
Language information
linguistic algorithm
language shares
Latin American
p library activities
library applications
library card catalog conversion
library card catalogs
library catalog card contents
library catalog operation
p library catalogs
library characteristics
p library collections
library community
library context
library cost structure
library expenditures
library explosion
library facilities
library file operations
library files
library holdings
Library Management Tool
library market
library materials
library mechanization
Library of Congress acquisition data

Library of Congress acquisition shares
Library of Congress classification system
Library of Congress nonserial acquisitions
Library of Congress size distribution
p library operations
library personnel
library procedures
library services
library shelf lists
library structure
p library systems
librarylike activities
line printers
linear string
lines per second
linguistic biases
linguistic constructions
linguistic data-objects
linguistic exploitation
linguistic records
linguistic partitions
linguistic subpopulation shares
list structure
literate population
load requirements
location information
log graph paper
logarithmic graph paper
p logarithmic scales
Logical operations
loss rate
"lower level" languages
p machine-readable catalogs
machine-readable library catalogs
machine-readable materials
machine-readable subject authority lists
machine change
machine design
machine elements
machine inquiries
p machine languages
machine language instructions
machine methods
machine output
machine rules
machine time usage

machine use
magnetic cores
magnetic discs
magnetic drums
p magnetic tapes
main file
management tool
manipulative operations
manpower costs
manual generation
manual operations
manual strategies
manuscript form
map classification category
marginal improvements
mathematical computation
mathematical exercise
Mathematical Journal Titles
mathematics
mathematics faculty committee
mathematics journals
mean growth
mechanical errors
mechanical translation
mechanization context
methodological principle
i Misspelled words
model cost equations
monetary inflation
monograph collection
monographic letter frequencies
i month portion
multi-language manipulation procedures
multicharacter vowel string
multiple copy graphic arts quality author list
musical scores
national accounts statistics
national economic growth
national economy
national origins
national publications
national publishers
national statistcial data collection processes
natural languages
(new) acquisitions information
non-English words
non-numerical procedures
nonlibrary customers
nonlinear scales

nonoriental monograph acquisitions
nonpamphlet items
nonserial Fondren sample
nonserial shelf list cards
nonserial textual works
nonstationary growth periods
nonstationary intervals of
library growth
non stationary time series
"normal probability paper"
normal distribution
normal probability distribution
normative measures
number field
numeric symbols
numerical computation
off-site areas
on-line input
open-stack libraries
optical character recognition
equipment
optional parameters
order-of-magnitude changes
order date
order file records
(order of) magnitude cost reductions
(order of) magnitude cost variations
(order of) magnitude decisions
(order of) magnitude gains
order operation
order system
order system file
order system reports
ordinal numbers
out-of-date catalog
output error signals
output list
output machines
output printers
output sheet
page counts
page design
paper costs
(paper) tape input
parallel search logic
pattern-matching facilities
pattern-valued functions
pattern primitives
per capita growth

per unit basis
percentage growth
personal author
personal authorship
p personal incomes
photo-offset reproduction
(physical) volumes per serial title
pilot study
plant expansion
political disintegration
political issues
population growth
potential control
print runs
printed copies
(printing and) binding costs
printing cost
i Printing Type Faces
private endowment funds
Private Finance
probability scale graph
process flow
x processing bibliographic records
x processing linguistic information
production costs
production economies
production processes
productivity per dollar
profound machine language level study
program errors
program routines
proper-name entries
i Proper names
proper scale compression
propositional calculus"
public acceptance
public card catalog
public catalog losses
p public libraries
public sales
public use
publication cost
i publication field
publication growth
publishing industries
punch paper tape
quality performance figures
quality point
quantal jump characteristics
quantal jumps
i random access

p random samples
record entry
recursive processes
refugee movements
relative frequencies
relative frequecy distribution
relative merit
relative performance
relative significance
relative size
reliable data
rental figure
report system
x reprogramming costs
research effort
research grants
research purposes
retrieval processes
retrieval requests
p retrospective files
p retrospective materials
run costs
salary structure
sample cards
school cooperation
scientific effort
scientific machines
scientific periodical literature
scientific publication
scientific research
selection criteria
selection operation
selection procedure
selection processes
selection technique
semibold type faces
semilogarithmic paper
p serial publications
serials shelf list
service bureaus
set theoretic operations
share distribution
shelf list circulation file
Shelf List Statistics
shelf space
significant acquisitions - GNP disagreement
social dislocation
social ideologies
social phenomena
social systems

sort operation
source statement language
source statement structure
special purpose bibliographies
e square inch
staff expansion requests
standard algebraic form
standard algebraic notation
standard precedence conventions
stationary growth rate
statistical correlation theory
x statistical distribution
x statistical distributions
statistical ensemble
statistical indicator
statistical relationship
statistical summary
statistical uniformities
status indicators
storage media
storage space
string contents
string processor
i structure
subject-oriented bibliographies
Subject-Title catalog
x subject area
x subject areas
x subject bibliographies
x subject bibliography
subject catalog volume
subject classes
subject coverage
subject definition
subject designation
subject material
subject matter
subject volumes
subject words
suburban population
summary information
supervision costs
symbol strings
Symbolic Expressions
systematic way
tape costs
technological advances
telephone companies
p telephone directories
text samples

textual works
"third generation" computers
time advantage
(time and) motion studies
time benefits
time constraints
Time Field
time information
time scale
time schedule
time variation
title-word access
title-word information
x title card
x Title cards
title indices
title languages
transcription errors
transliteration schemes
tray contents
trend curve
turn-around times
type face catalog
type face size
p type faces
type fonts
type size
type styles
undergraduate library
undergraduate student
Union catalogs
x unit cost
x unit costs
p university libraries
university library book catalog
university library systems
university order staff
university rate
usage rates
user-library complex
"user codes"
User cost
(user) cost factors
i utility
utilization costs
vertical scale
e XYZ Library
yield rate per item

does not characterize indexible sequences. 'x' indicates that a manual error has been made; in some cases a verb gerund has not been deleted in the stop list step of the algorithm, so a sequence appears in the later stages of the algorithm when it ought to have been deleted at the first stage. For example, the sequence *processing bibliographic records* contains the structural stop sequence *-ing* indicating the gerund form; exclusion of this word at the stop list stage would have left the subsequence *bibliographic records* for consideration, which appears in the index anyway because it occurred in more than one additional location. The indicator 'e' means that the sequence has been excluded by the human posteditor. Two such sequences are noted: *square inch*, which should perhaps inhabit the stop list, and *XYZ Library*, which must be considered because one of the special format inclusion conditions is that sequences containing all capitalized words are indexed regardless of the number of text locations to which they refer; but this instance doesn't supply any useful information. It is a stylistic curiosity. Finally, certain sequences in the table are preceded by a parenthesized word. For instance, *(first) generation machines* appears. The algorithm generated *generation machines*; the preceding text word was included in the list to help the reader to understand the context of the sequence, which, following the algorithm, was excluded from the index.

Quantitatively, this algorithmic index is not significantly different from the manually produced indexes analyzed in Chapter IV. The gross size of the index is 5 pages as compared to 157 pages of text, a text to index ratio of 31.4 to 1. The index entry length distribution is given in Table 5.3.

Table 5.3

Index Entry Length Distribution
Computerized Library Catalogs

Number of Words	Frequency	Cumulative Frequency
1	82	82
2	190	272
3	44	316
4	13	329
5	6	335
6	4	339
7	1	340

The percentage of one-word entries (24%) is higher than the average number of one-word entries in the subsample from the Fondren Index Sample (13%). Although this is not a significant deviation (more than 17% of the indexes in the subsample had more than 24% one-word entries) it is worthy of some comment: the basic algorithm suppresses one-word entries, with exceptions. In this case the exception rule was to include capitalized one word entries. Thus, even though the algorithm is designed to operate against one-word entries, the proportion occurring is still on the high side.

The distribution of entries by number of words is shown in Figure 5.3. The distribution is reasonably approximated by the lognormal distribution. The arithmetic mean of the distribution is 2.08 words per entry, compared to 3.68 words per entry for the subsample as a whole. Although there is again some cause to question whether this is a significant deviation, there is an underlying weakness in the form of the algorithm as it was used in this example. The algorithm excludes entries of the word X of Y. In (4) the structure-word-free entries were found to have a mean number of words per entry of 2.12, almost exactly the average found for this algorithmic index. However, the structure-word-free entries of (4) made up only 55% of the total number of entries. In Chapter VI we shall return to this question in analyzing the output of the basic algorithm where the capability to generate entries of the form X of Y has been included.

The absence of structure-word entries also tends to depress the overall size of the index. Although the bulk size, measured in pages is approximately 1/30th of the text size (as would be expected), the ratio of bulk of the index to bulk of the text measured in number of characters is approximately half of this figure. (Not only are the index entries somewhat shorter than would be found in the manual indexes, the text density is approximately 3,150 characters per page as compared to the mean of 2,400 characters per page.)

The lack of structure-word entries also tends to distort the page location distribution, (Table 5.4).

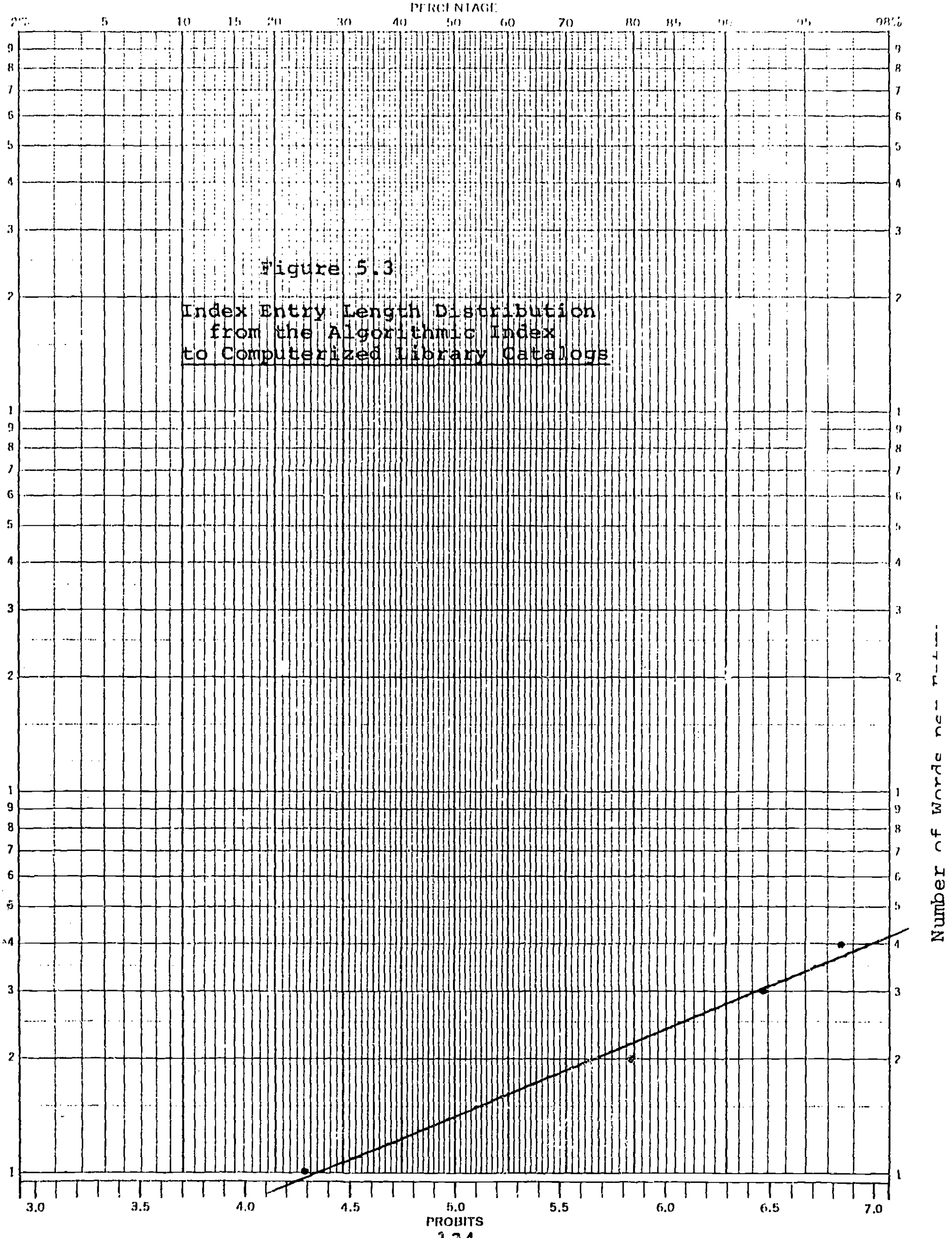


Figure 5.3

Index Entry Length Distribution from the Algorithmic Index to Computerized Library Catalogs

Table 5.4

INDEX PAGE LOCATION DISTRIBUTION
COMPUTERIZED LIBRARY CATALOGS

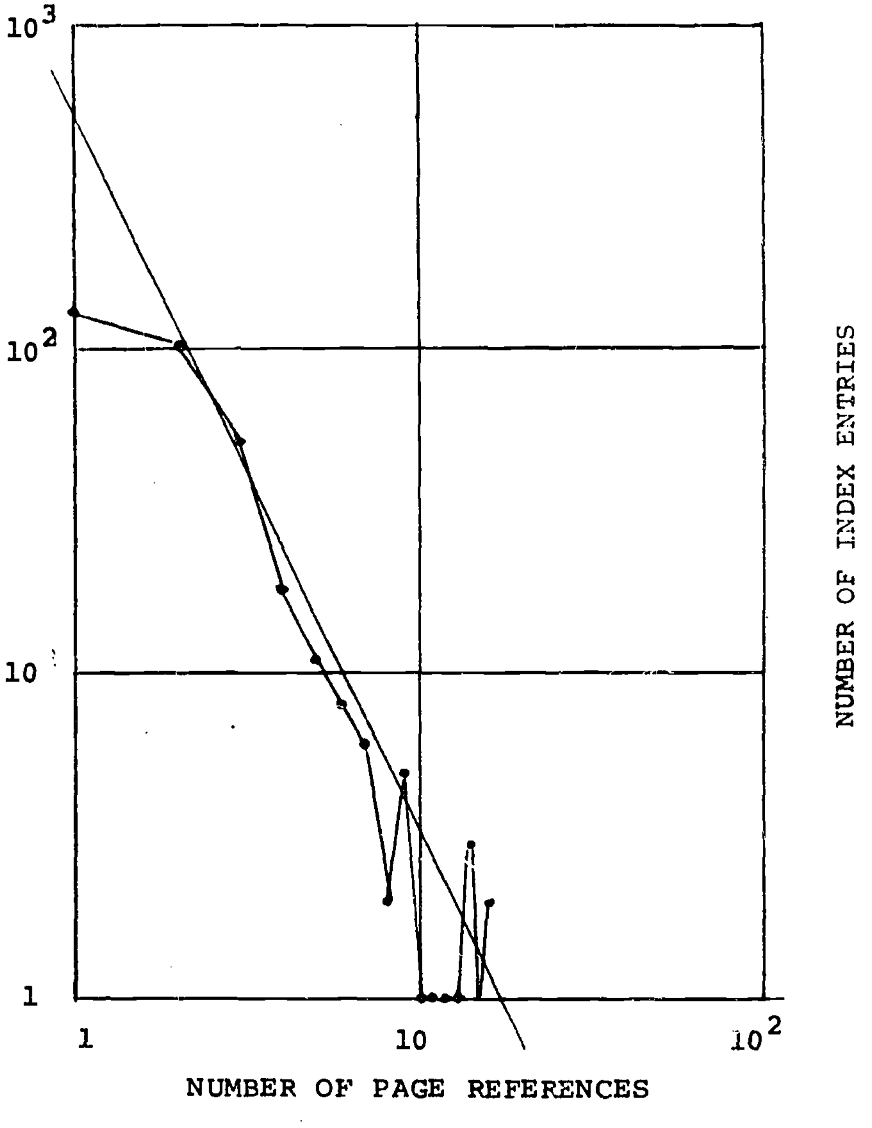
Number of Page Locations Per Entry	Frequency	Cumulative Frequency
1	127	127
2	102	229
3	52	281
4	18	299
5	11	310
6	8	318
7	6	324
8	2	326
9	5	331
10	1	332
11	1	333
12	1	334
13	1	335
14	3	338
15	0	338
16	2	340

The graph of the index page location distribution is shown in Figure 5.4. Here it is evident that the number of entry with but a single page location is significantly lower than the overall trend line for the rest of the data. Further, the bend in the data occasioned by this low value is sharper than for any of the distributions in the subsample from the Fondren Index Sample (see Appendix II). Interconnection of the entries with structure words would clearly tend to break apart entries presently agglomerated, thus reducing the number of multiply occurring entries. Ignoring the low number of singly occurring entries, the Zipf-Mandelbrot slope is 2.17, well within the range of values found for the manually produced indexes.

The arithmetic mean of the number of page locations per entry is 3.19, nearly double the figure found for the subsample of the Fondren Index Sample. However, this value is distorted by the fact that consecutive page locations were not agglomerated into single locations as is normally done in manual indexing. When this factor is corrected, the average number of page locations per entry becomes 2.14. As this value would be further reduced by inclusion of structure-word entries, it would appear that this variation is not at all significant.

Figure 5.4

Index Page Location Distribution
from the Index to
Computerized Library Catalogs



In sum, aside from the failure to include structure-word entries or to agglomerate consecutive page locations, the statistical shape of the algorithmic index to *Computerized Library Catalogs* appears sound. This is not to say that the index is entirely comparable to a manually produced index. However, the first requirement in automating a process traditionally done manually is to meet the basic size constraints. Further developments in the technique will be illustrated in the next chapter to demonstrate that even closer approximations are possible.

References

1. Salton, Gerard, *Automatic Information Organization and Retrieval*, McGraw-Hill Book Co., New York, 1968.

2. Meetham, A. R. and R. A. Hudson, editors, Encyclopaedia of Linguistics, Information and Control, Pergamon Press, Oxford, 1969.

3. Dolby, J. L., V. Forsyth, and H. L. Renikoff, *Computerized Library Catalogs: Their Growth, Cost and Utility*, the M. .T. Press, Cambridge, 1969.

4. Dolby, J. L., "The Structure of Indexing: the Distribution of Structure-Word-Free Back-of-the-Book Entries", *Proceedings of the American Society of Information Science*, *5* (1968), 65-72.

5. Dolby, J. L. and W. E. Houchin, *A Modular Suite of Programs for System ABC*, R & D Consultants Co., Los Altos, California, 1969.

6. Dolby, J. L., W. E. Houchin, H. L. Resnikoff, and Roger Stark, *Non-Numeric Programming Language Studies: ALTEXT II.*, Final Report to the U. S. Air Force Office of Scientific Research, Contract #F44620-69-C-0094, R & D Consultants Co., Los Altos, California, 1970.

CHAPTER VI

AMALGAMATIVE ACCESS MECHANISMS

AMALGAMATIVE ACCESS MECHANISMS

INTRODUCTION

The model proposed in Chapter 2 shows that the search for access mechanisms must be conducted in compressive powers of 30. It is principally the *relative size* of an access mechanism that determines its utility. That a compression of 30 must be effected in order to move from one access level to the next, and that the boundary between access levels corresponds to compression of about a factor of 5 implies that there cannot be very many possible access mechanisms to a particular level of information storage. For instance, if the level to be accessed is the book, then one must ask what natural subsets of information there are in a book which constitute about one-thirtieth of it. As has already been pointed out, the average index to the average book compresses the text by a factor of 31.8, so the book index is a viable access mechanism. Studies of abstracts of papers appearing in mathematical journals show that the average complete abstract produces a compression of about 30.6, so the journal paper abstract is also a viable access mechanism. The *book abstract* should require about $276.6/(2e)^2 = 9.3$ pages; we do not have reliable information about the average length of book reviews in the professional literature, but this appears to us to be a possible mean for scholarly reviews. On the other hand, the capsule reviews of popular books that appear in newspapers and other popular media, and in some scholarly publications, are much shorter--perhaps the equivalent of one or two pages--and lie on the boundary between the levels of access mechanisms to books and access mechanisms to access mechanisms to books, the latter operating at the level of an enlarged table of contents such as regularly appeared in previous centuries, and still sometimes do, *viz.*, Hans Zinsser's *Rats, Lice and History*'s table of contents from which we extract the following:

I. In the nature of an explanation and an apology

II. Being a discussion of the relationship between science and art

III. Leading up to the definition of bacteria and other parasites, and digressing briefly into the question of the origin of life

IV. On parasitism in general, and on the necessity of considering the changing nature of infectious diseases in the historical study of peidemics

V. Being a continuation of Chapter IV, but dealing more particularly with so-called new diseases and with some that have disappeared.

and so forth.

Another way of looking at the problem of discovering possible methods for accessing books is this: the number of characters in a book is about $(2e)^8$; reduction of a factor of $(2e)^2$ leads to an information store about the size of the index; further reduction by a factor of $(2e)^2$ to the next access level leads to a store of the size of the table of contents. Another reduction by the same factor produces $(2e)^2 \cong 30$ characters, which is nearly the size of a book title, as we have determined in a preliminary fashion from a small uniform subsample of the Fondren Sample. In fact, that estimate was 34.2 characters for monographs in the sample regardless of language of title; had the subsample been restricted to English language titles, the average length would have been shorter. A final usable reduction is effected by another division by $(2e)^2$, leading to one character access mechanism such as that provided by the Library of Congress one letter class designation.

The important point is that every access level is filled. Further study of possible new access mechanisms must therefore be constrained to access mechanisms of the same size as those that already exist. A natural question that arises is whether it is desirable to have two access mechanisms of the same size for a particular information system. That such duplication does already exist is easy to demonstrate:

1. The Author, Title, and Shelf orderings or a library card catalog are all essentially of the same size: roughly, one card image for each title in the collection. (The subject heading ordering is generally slightly larger, but still at the same access level as the others.)

2. The table of contents for a book is at the same access level as the catalog record.

3. Abstracts to journal articles appear in abstract journals as well as the index entries that are frequently published at the end of the year in the journal. Both of these access mechanisms are first order devices.

and of course other examples involving titles, descriptors, etc. can easily be found.

Thus the size of an access mechanism, though it is of first importance in describing the nature of the access it provides, is not sufficient to completely describe its characteristics. A second consideration that must be taken into account is easily illustrated by considering the sequences:

Article, Abstract, Title

and

Book, Index, Table of Contents

In the first sequence, each access device is acting simply to compress the contents of the primary information store. In the second sequence, each access mechanism is itself a set of lower order access mechanisms collected and sorted in a useful ordering. The abstract and the title provide the user with the opportunity to determine whether the document so described is likely to be relevant to his need for information, in a general way. The index and the table of contents provide the user with information about the contents of the document together with the location of particular pieces of information in the document.

The crucial question is that of agglomeration: an index is an agglomeration of entries; a table of contents is an agglomeration of entries; on the other hand both title and the abstract are entities themselves rather than being agglomerations of other entities. It seems clear from what has gone before that the minimal unit for agglomeration is the first level unit (about 30 characters). Thus both the table of contents and the index are agglomerations of first level units. However, higher level agglomerations exist: the abstract journal is an agglomeration of second level units, as is a publication devoted to the republication of the tables of contents of journals. Although we have not yet completed our study of dictionaries and encylopedias, it is clear that each of these important access devices are agglomerations of higher level entities.

In this sense, an access mechanism can be described first by its total size and secondly by the size of the primary entries that it agglomerates. Thus an abstract is zero level agglomeration of second level entries; a table of contents is a first level agglomeration of first level entries; and an index (to a book) is a second level agglomeration of first level entries.

There are at least two other factors that must be taken into account: a cumulative index to a series of books on statistics obviously plays a different role than the index to an encyclopedia even though both are third level agglomerations of first level entries. The difference here is that the encyclopedia is itself an agglomeration of second or higher level access mechanisms, while the books are primary information stores. The difference in these two mechanisms would almost undoubtedly show up in the slope (in the Mandelbrot sense discussed earlier) of the index.

Finally, there are access mechanisms clearly dedicated to "non-subject" access, e.g., author indexes, list of publications by publisher, place of publication, time of publication, etc. which play a major role in library access systems.

Consider a collection of titles--such as book titles--of items which corpass a range of subject matter. The card catalog title list is one ordering of such a collection. If the collection is reordered to bring together all titles which contain a given information bearing word, then access to the collection is significantly increased.

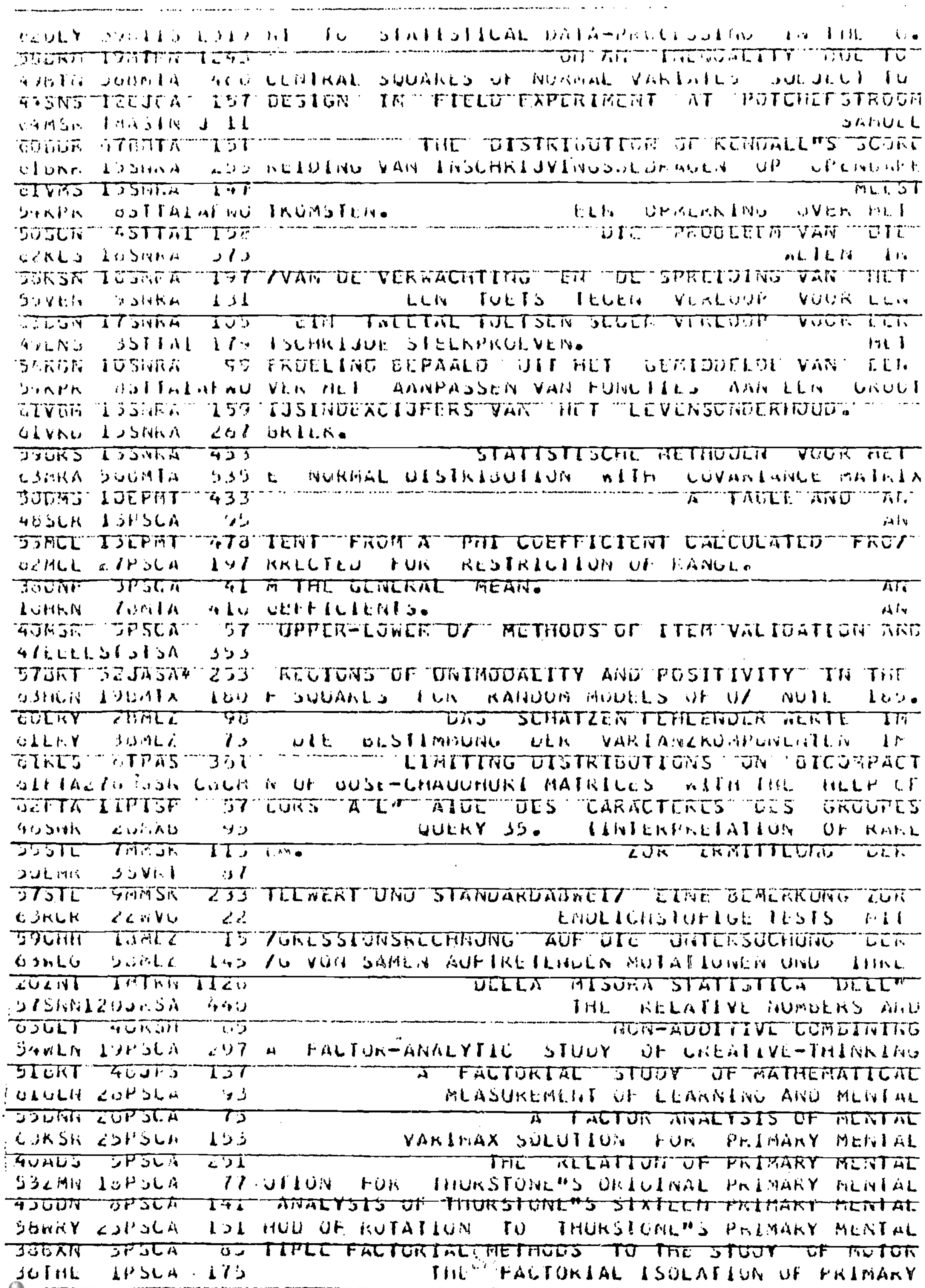
Studies of such access mechanisms have been underway for some time, although none of them are generally available. One of the most advanced title access mechanisms is that prepared at Princeton University under the direction of J. W. Tukey; it is a sophisticated permuted title index consisting of more than 25,000 titles of journal papers in the field of statistics. Since the average length of a paper in mathematics is about 13.8 (normalized) pages, a title represents a compression of about two access levels, for the title as it appears in a permuted title index carries information about the journal and author as well, requiring about 130 characters. A sample page from the Princeton permuted title index is shown in Figure 6.1.

General considerations suggest that a permuted title list of book titles for the Library of Congress letter class subcollections of archival libraries would be a useful tool, and one which would be readily obtainable as a byproduct of the existence of a machinable catalog data base.

Another type of amalgamative access mechanism, which provides access to a collection of items belonging to the same access level rather than to only one item can be constructed by performing the process normally used to construct a standard access mechanism on the output produced by another. For instance, we have studied the utility of indexing abstracts to journal papers in the statistical literature. The abstracts are normally provided with the papers; they have been converted to machinable form and an elementary version of the indexing algorithm described in Chapter 5 was applied to them. Appendix A5 exhibits the abstracts to 50 papers, the associated abstract indexes produced by application of the algorithm, and a cumulative list of the resulting index terms with references to the articles in which the terms appeared. We reproduce an abstract with its index as Figure 6.2 and a page from the cumulative abstract index as Figure 6.3. The abstract index was the first processed in this series; it is perhaps not entirely typical of the output from the algorithm. We have also processed the same data using a variant of the algorithm which ignores in its analysis stage the presence of the preposition "of" and consequently will produce index entries like "basic limit theorem of renewal theory" which appears in Figure 6.2 only by way of its constituent phrases "basic limit theorem" and "renewal theory".

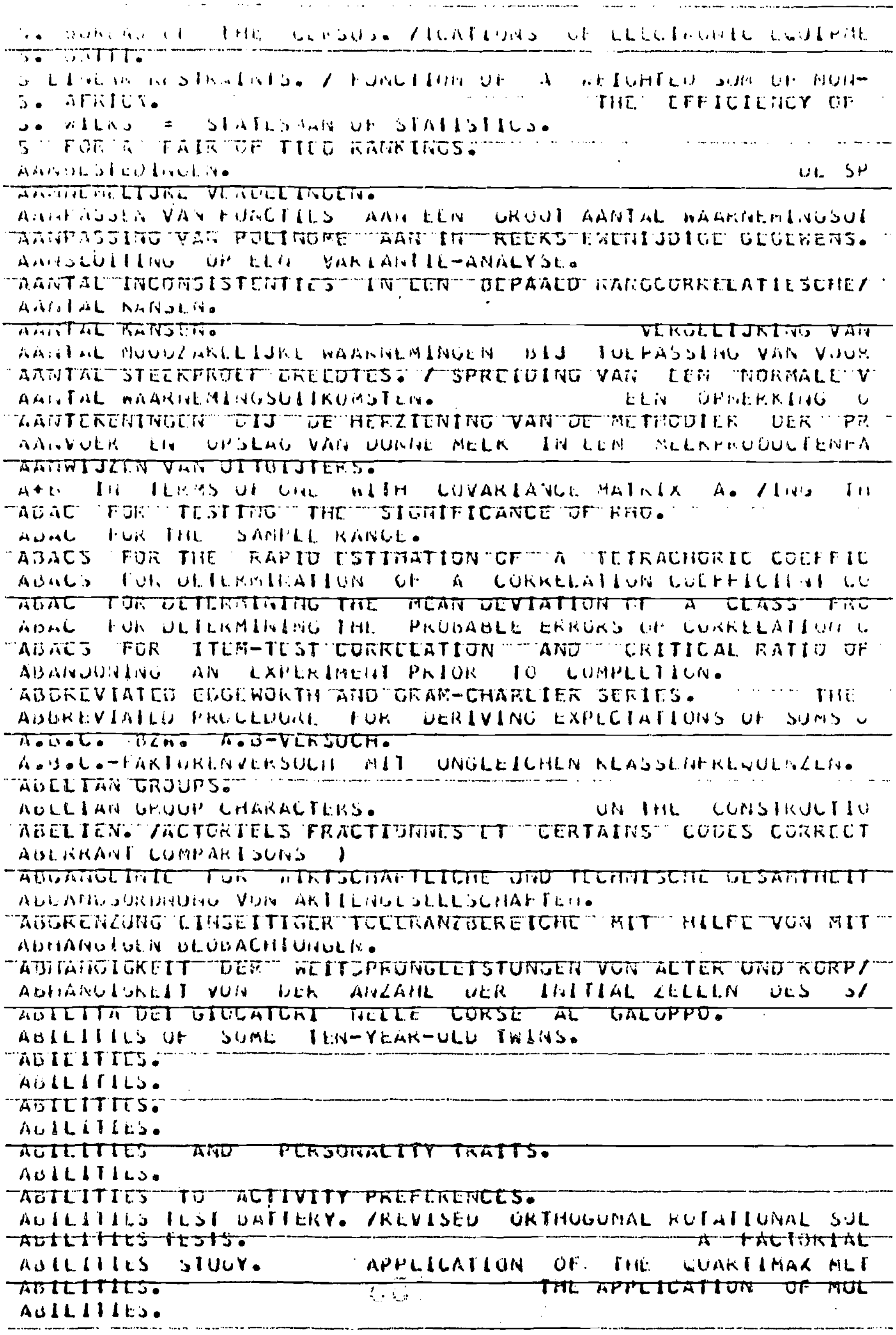
An index to an abstract is a hybrid form of access mechanism. The abstract already contains a large proportion of significant phrases which are repeated in the extractive output of the indexing algorithm. There is therefore no hope that an index to an abstract can provide a compression of a full factor of 30 that would be necessary to descend from one access level to that immediately below it. In fact it appears that indexing abstracts leads to a compression of about 15; since this is significantly greater than (2e), such a procedure does

Figure 6.1
Permuted Title Index Page (Left Hand Side)



```
62DLY 59C115 L517 NT  TO  STATISTICAL DATA-PROCESSING  IN THE  U.
55GRH 17MTFR 1295                      ON AN  INEQUALITY  DUE TO
49BTG 50BMIA  400 CENTRAL SQUARES OF NORMAL VARIATES  SUBJECT TO
49SNS 12EJCA  157 DESIGN  IN  FIELD EXPERIMENT  AT  POTCHEFSTROOM
64MSK 1MASIN J 11                                         SAMUEL
60GUR 47BMTA  151          THE  DISTRIBUTION OF KENDALL"S SCORE
61DKR 15SNRA  255 REIDING VAN INSCHRIJVINGSBEDRAGEN  OP  OPENBARE
61VMS 15SNRA  147                                          MEEST
54KPR  65TTAIAFWO TROMSTEN.               EEN  OPMERKING  OVER HET
50SCN  45TTAI 192                         DIE  PROBLEEM VAN  DIE
62KLS 16SNRA  575                                       AEIEN  IN
53KSN 10SNRA  197 /VAN DE VERWACHTING  EN  DE SPREIDING VAN  HET
59VER  9SNRA  131          EEN  TOETS  TEGEN  VERLOOP  VOOR EEN
53DGN 17SNRA  105   EEN  TWEETAL TOETSEN TEGEN VERLOOP  VOOR EEN
49ENS  35TTAI 179 ISCHRIJDE STEEKPROEVEN.                    HET
54RGN 10SNRA   59 EROELING BEPAALD  UIT HET  GEMIDDELDE VAN  EEN
54KPR  65TTAIAFWO VER HET  AANPASSEN VAN FUNCTIES  AAN EEN  GROOT
61VGM 15SNRA  159 IJSINDEXCIJFERS VAN  HET  LEVENSONDERHOUD.
61VKD 15SNRA  267 BRIEK.
59GRS 15SNRA  433                STATISTISCHE METHODEN  VOOR HET
63MRA 50BMIA  535 E  NORMAL DISTRIBUTION  WITH  COVARIANCE MATRIX
50DMS 10EPMT  433                                A  TABLE AND  AN
48SCR 13PSCA   95                                             AN
53MCL 13EPMT  478 TENT  FROM A  PHI COEFFICIENT CALCULATED  FRO/
62MCL 27PSCA  197 RRECTED  FOR  RESTRICTION OF RANGE.
38DNF  3PSCA   41 M THE GENERAL  MEAN.                         AN
10HRN  70MIA  410 OEFFICIENTS.                                 AN
43MSR  8PSCA   57  UPPER-LOWER 0/  METHODS OF ITEM VALIDATION AND
47EEELSESISA  355
57BRT 52JASA4 253  REGIONS OF UNIMODALITY AND POSITIVITY  IN THE
63HCN 19BMIX  180 F SQUARES  FOR  RANDOM MODELS OF U/  NOTE  165.
60ERY  28MLZ   98               DAS  SCHATZEN FEHLENDER WERTE  IM
61ERY  30MLZ   75   DIE  BESTIMMUNG  DER  VARIANZKOMPONENTEN  IM
61KLS  6TPAS  351             LIMITING DISTRIBUTIONS  ON  BICOMPACT
61TTA26 3SK CSCH N OF BOSE-CHAUDHURI MATRICES  WITH THE  HELP OF
62FTA 11PISF   57 CORS  A L"  AIDE  DES  CARACTERES  DES  GROUPES
40SNR  20MAB   95            QUERY 35.  (INTERPRETATION  OF RANK
55STL  7MMSK  115 EN.                           ZUR  ERMITTLUNG  DER
50EMR  35VRT   87
57STL  9MMSK  233 TELWERT UND STANDARDABWEIZ  EINE BEMERKUNG ZUR
63RCR  22WVG   22                      ENDLICHSTUFIGE TESTS  MIT
59GHR  13MLZ   15 /GRESSIONSRECHNUNG  AUF DIE  UNTERSUCHUNG  DER
63WLG  58MLZ  145 /G VON SAMEN AUFTRETENDEN MUTATIONEN UND  IHRE
20ZNI  1MTRN 1120                  DELLA  MISURA STATISTICA  DELL"
57SRN1205ESA  440                        THE  RELATIVE NUMBERS AND
65GLT  40RSM   65                            NON-ADDITIVE COMBINING
54WLN 19PSCA  297 A  FACTOR-ANALYTIC  STUDY  OF CREATIVE-THINKING
51CRT  40JPS  137          A  FACTORIAL  STUDY  OF MATHEMATICAL
61GLR 26PSCA   93             MEASUREMENT OF LEARNING AND MENTAL
55DRR 20PSCA   75                   A  FACTOR ANALYSIS OF MENTAL
63KSR 25PSCA  153         VARIMAX SOLUTION  FOR  PRIMARY MENTAL
40ADS  5PSCA  251                THE  RELATION OF PRIMARY MENTAL
53ZMN 18PSCA   77 OTION  FOR  THURSTONE"S ORIGINAL PRIMARY MENTAL
45DDN  8PSCA  141  ANALYSIS OF THURSTONE"S SIXTEEN PRIMARY MENTAL
58WRY 23PSCA  151 HOD OF ROTATION  TO  THURSTONE"S PRIMARY MENTAL
38BXN  3PSCA   83 TIPLE FACTORIAL METHODS  TO THE STUDY  OF MOTOR
36THE  1PSCA  175            THE  FACTORIAL ISOLATION OF PRIMARY
```

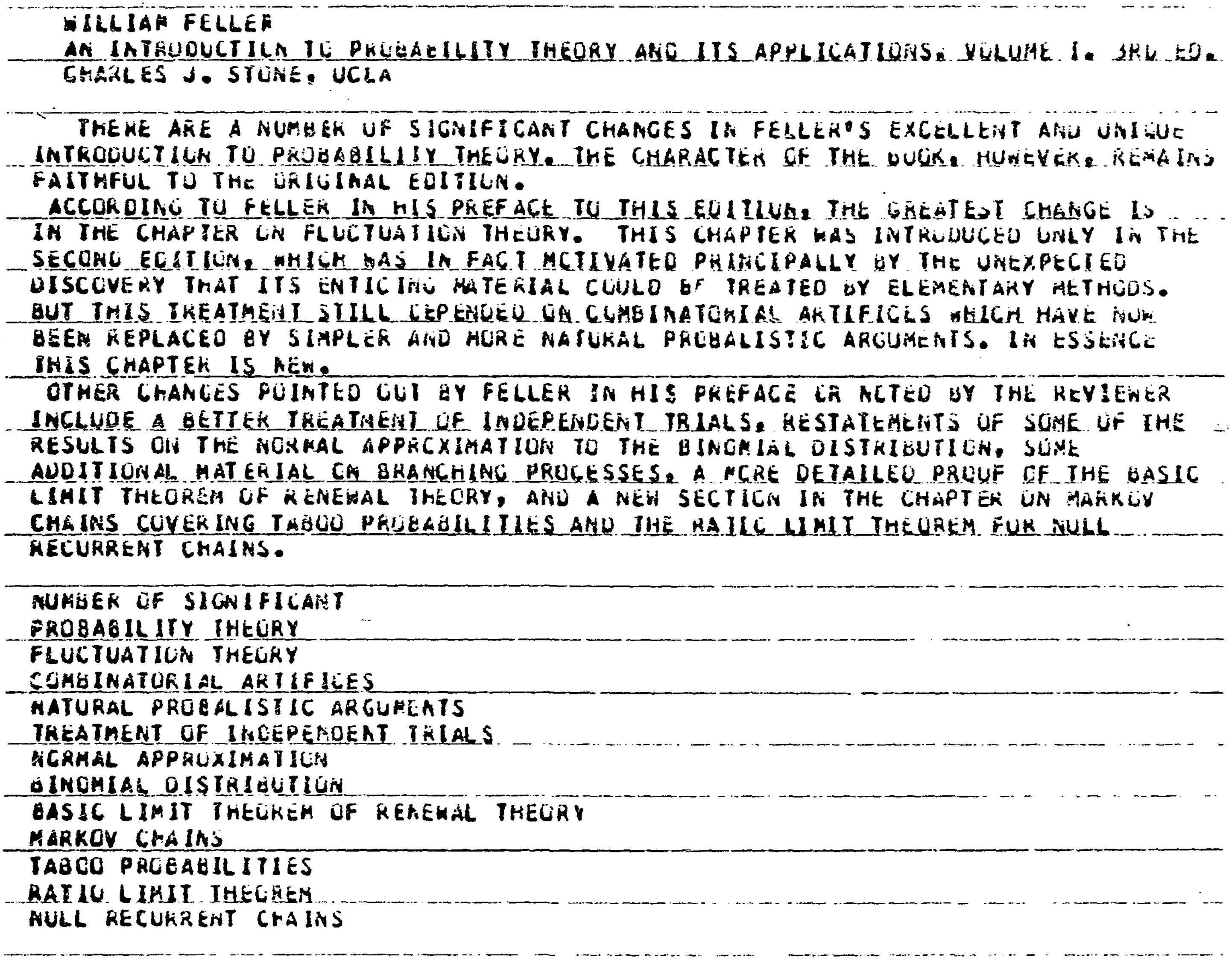
Figure 6.1
Permuted Title Index Page (Right Hand Side)



```
S. BUREAU OF THE CENSUS. /ICATIONS OF ELECTRONIC EQUIPME
S. SMITH.
S LINEAR RESTRAINTS. / FUNCTION OF A WEIGHTED SUM OF NON-
S. AFRICA.                                 THE EFFICIENCY OF
S. WILKS = STATESMAN OF STATISTICS.
S FOR A PAIR OF TIED RANKINGS.
AANBESTEDINGEN.                                         DE SP
AANNEMELIJKE VERDELINGEN.
AANPASSEN VAN FUNCTIES AAN EEN GROOT AANTAL WAARNEMINGSUI
AANPASSING VAN POLINOME AAN IN REEKS EWENIJDIGE GEGEWENS.
AANSLUITING OP EEN VARIANTIE-ANALYSE.
AANTAL INCONSISTENTIES IN EEN BEPAALD RANGCORRELATIESCHE/
AANTAL KANSEN.
AANTAL KANSEN.                                 VERGELIJKING VAN
AANTAL NOODZAKELIJKE WAARNEMINGEN BIJ TOEPASSING VAN VOOR
AANTAL STEEKPROEF GROOTES. / SPREIDING VAN EEN NORMALE V
AANTAL WAARNEMINGSUITKOMSTEN.                EEN OPMERKING O
AANTEKENINGEN BIJ DE HERZIENING VAN DE METHODIEK DER PR
AANVOER EN OPSLAG VAN DUNNE MELK IN EEN MELKPRODUCTENFA
AANWIJZEN VAN UITBIJTERS.
A+B IN TERMS OF ONE WITH COVARIANCE MATRIX A. /ING IN
ABAC FOR TESTING THE SIGNIFICANCE OF RHO.
ABAC FOR THE SAMPLE RANGE.
ABACS FOR THE RAPID ESTIMATION OF A TETRACHORIC COEFFIC
ABACS FOR DETERMINATION OF A CORRELATION COEFFICIENT CO
ABAC FOR DETERMINING THE MEAN DEVIATION OF A CLASS FRC
ABAC FOR DETERMINING THE PROBABLE ERRORS OF CORRELATION C
ABACS FOR ITEM-TEST CORRELATION AND CRITICAL RATIO OF
ABANDONING AN EXPERIMENT PRIOR TO COMPLETION.
ABBREVIATED EDGEWORTH AND GRAM-CHARLIER SERIES.       THE
ABBREVIATED PROCEDURE FOR DERIVING EXPECTATIONS OF SUMS O
A.B.C. BZW. A.B-VERSUCH.
A.B.C.-FAKTORENVERSUCH MIT UNGLEICHEN KLASSENFREQUENZEN.
ABELIAN GROUPS.
ABELIAN GROUP CHARACTERS.                    ON THE CONSTRUCTIO
ABELIEN. /ACTORIELS FRACTIONNES ET CERTAINS CODES CORRECT
ABERRANT COMPARISONS )
ABGANGLINIE FUR WIRTSCHAFTLICHE UND TECHNISCHE GESAMTHEIT
ABGANGSORDNUNG VON AKTIENGESELLSCHAFTEN.
ABGRENZUNG EINSEITIGER TOLERANZBEREICHE MIT HILFE VON MIT
ABHANGIGEN BEOBACHTUNGEN.
ABHANGIGKEIT DER WEITSPRUNGLEISTUNGEN VON ALTER UND KORP/
ABHANGIGKEIT VON DER ANZAHL DER INITIAL ZELLEN DES S/
ABILITA DEI GIOCATORI NELLE CORSE AL GALOPPO.
ABILITIES OF SOME TEN-YEAR-OLD TWINS.
ABILITIES.
ABILITIES.
ABILITIES.
ABILITIES.
ABILITIES AND PERSONALITY TRAITS.
ABILITIES.
ABILITIES TO ACTIVITY PREFERENCES.
ABILITIES TEST BATTERY. /REVISED ORTHOGONAL ROTATIONAL SOL
ABILITIES TESTS.                                 A FACTORIAL
ABILITIES STUDY.        APPLICATION OF THE QUARTIMAX MET
ABILITIES.                             THE APPLICATION OF MUL
ABILITIES.
```

Figure 6.2

Abstract and Abstract Index



WILLIAM FELLER
AN INTRODUCTION TO PROBABILITY THEORY AND ITS APPLICATIONS. VOLUME I. 3RD ED.
CHARLES J. STONE, UCLA

THERE ARE A NUMBER OF SIGNIFICANT CHANGES IN FELLER'S EXCELLENT AND UNIQUE INTRODUCTION TO PROBABILITY THEORY. THE CHARACTER OF THE BOOK, HOWEVER, REMAINS FAITHFUL TO THE ORIGINAL EDITION.

ACCORDING TO FELLER IN HIS PREFACE TO THIS EDITION, THE GREATEST CHANGE IS IN THE CHAPTER ON FLUCTUATION THEORY. THIS CHAPTER WAS INTRODUCED ONLY IN THE SECOND EDITION, WHICH WAS IN FACT MOTIVATED PRINCIPALLY BY THE UNEXPECTED DISCOVERY THAT ITS ENTICING MATERIAL COULD BF TREATED BY ELEMENTARY METHODS. BUT THIS TREATMENT STILL DEPENDED ON COMBINATORIAL ARTIFICES WHICH HAVE NOW BEEN REPLACED BY SIMPLER AND MORE NATURAL PROBALISTIC ARGUMENTS. IN ESSENCE THIS CHAPTER IS NEW.

OTHER CHANGES POINTED OUT BY FELLER IN HIS PREFACE OR NOTED BY THE REVIEWER INCLUDE A BETTER TREATMENT OF INDEPENDENT TRIALS, RESTATEMENTS OF SOME OF THE RESULTS ON THE NORMAL APPROXIMATION TO THE BINOMIAL DISTRIBUTION, SOME ADDITIONAL MATERIAL ON BRANCHING PROCESSES, A MORE DETAILED PROOF OF THE BASIC LIMIT THEOREM OF RENEWAL THEORY, AND A NEW SECTION IN THE CHAPTER ON MARKOV CHAINS COVERING TABOO PROBABILITIES AND THE RATIC LIMIT THEOREM FOR NULL RECURRENT CHAINS.

NUMBER OF SIGNIFICANT
PROBABILITY THEORY
FLUCTUATION THEORY
COMBINATORIAL ARTIFICES
NATURAL PROBALISTIC ARGUMENTS
TREATMENT OF INDEPENDENT TRIALS
NORMAL APPROXIMATION
BINOMIAL DISTRIBUTION
BASIC LIMIT THEOREM OF RENEWAL THEORY
MARKOV CHAINS
TABOO PROBABILITIES
RATIO LIMIT THEOREM
NULL RECURRENT CHAINS

Figure 6.3

Cumulative Index to 50 Abstracts (one page)

NONCENTRAL DISTRIBUTIONS	43
NONCENTRAL MULTIVARIATE BETA DISTRIBUTION	43
NONNULL DISTRIBUTIONS	29
NONPARAMETRIC ALTERNATIVE	42
NONPARAMETRIC ALTERNATIVES	37
NONPARAMETRIC TESTS	10
NORMAL ALTERNATIVES	37
NORMAL APPROXIMATION	1
NORMAL DISTRIBUTION	32
NORMAL DISTRIBUTION	33
NORMAL DISTRIBUTIONS	29
NORMAL DISTRIBUTIONS	48
NORMAL POPULATIONS	29
NORMAL THEORY LIKELIHOOD RATIO STATISTIC	27
NORMAL THEORY LIKELIHOOD RATIO TEST STATISTIC	27
NULL DISTRIBUTION	29
NULL DISTRIBUTIONS	25
NULL HYPOTHESIS	29
NULL HYPOTHESIS	29
NULL RECURRENT CHAINS	1
NUMBER SUCCESSES	23
NUMERICAL COMPARISONS	41
NUMERICAL EXAMPLES	29
ONE-DIMENSIONAL EMPIRICAL PROCESS CONVERGE	7
ONTO ITSELF	47
OPERATIONAL CHARACTERISTICS	17
OPTIMAL ALLOCATION	2
OPTIMAL ALLOCATION PROBLEMS	2
OPTIMAL STRATIFIED SAMPLING	2
OPTIMUM ALLOCATION	19
OPTIMUM BEST LINEAR	33
OPTIMUM BLUE,S	33
OPTIMUM NON-PARAMETRIC STATISTICS	30
ORDER ABSOLUTE CENTRAL MOMENT	26
ORDER STATISTICS	13
ORDER STATISTICS	33
ORDER STATISTICS	33
ORDER STATISTICS	33
OVERALL AVERAGE	40
P-DIMENSIONAL SPACE	11
P-DIMENSIONAL VARIATE	2
P-VARIATE NORMAL POPULATIONS	8
PAPER GENERALIZES	36
PAPER TREATS *-COMPARIS	17
PARAMETER SET	17
PARAMETER SETS	17
PARAMETRIC CLASSES	5
PAST OBSERVATIONS AVAILABLE	39
PHASE DISTRIBUTION	3
PHASE DISTRIBUTION	4
PHASE SERVICE TIME DISTRIBUTION	4
POPULATION MEAN	16
POPULATION MEAN	40
POPULATION SIZE	23

realize a compressive gain that may be useful for accessing the abstracts. It will certainly be useful for accessing the original documents when it is applied to a collection of abstracts and the resulting indexes are accumulated.

The page extracted from the middle of the cumulative abstract indexed reproduced as Figure 6.3 shows that one paper in the sample of 50 referred, via its abstract, to the "NON-CENTRAL MULTIVARIATE BETA DISTRIBUTION", and, since the abstract transmitted this phrase, the paper undoubtedly contains something of interest about this topic. Similarly note that eight papers referred to the "NORMAL" distribution in some form. The presence of spurious terms like "ONTO ITSELF" and "OPTIMUM BLUE'S" is no more than a minor annoyance in use of the index, and is of course due to inadequacies in the indexing algorithm's "stop list", which should certainly contain the word "ITSELF". There are other more subtle problems whose genesis is the indexing algorithm, but they are not so obtrusive as to mae the use of the list burdensome. For instance, the phrase "OPTIMUM BLUE'S occurs in the abstract, where it is defined to denote "OPTIMUM BEST LINEAR UNBIASED ESTIMATE"; this phrase certainly belongs in the index, but it is not clear that a user of the amalgammated index would recognize the technical meaning of "BLUE" until it had become a standard term of the field.

Indexing abstracts is of potential value in gaining access to the large numbers of journal papers which annually appear in the literature; coupled with permuted title access mechanisms, the abstract index should provide a rapid and reliable means of surveying the key content areas of papers without the time-consuming process of reading abstracts, which often limits one to a relatively narrow and current range of documents.

When compared to the earlier manual implementation of the algorithm on Computerized Library Catalogs, the machine implementation of the algorithm differs in several ways, aside from the obvious fact that the machine is entirely consistent in its application where manual procedures cannot be. The raw data for the machine test on the statistical abstracts was keypunched in all upper case, as a matter of convenience. Hence, the rule to keep capitalized one-word entries was inoperative in this run. Further, no attempt has been made to include see or see-also types references in the machine implementation. On the other hand, the machine implementation includes logic to allow structure-word entries where the manual implementation did not.

These differences are reflected in the statistics describing the entry length and page location distributions.

Table 6.1 provides the entry length (in number of words) distribution for the machine index to the statistical abstracts.

Table 6.1

Entry Length Distribution
Algorithmic Index to Statistical
Abstracts

Number of Words	Frequency	Cumulative Frequency
1	0	0
2	315	315
3	233	548
4	125	673
5	54	727
6	19	746
7	6	752
8	2	754

Comparing this distribution to the comparable distribution for the manually implemented algorithmic index to Computerized Library Catalogs (Table 5.3) one sees that the proportion of one-word-entries has been reduced to zero (because there is no logic available to permit one-word-entries) and that the overall average entry length has been increased from 2.08 words per entry to 3.01 words per entry. The main factor in this increase is the introduction of structure-word entries, although the absence of one-word-entries has a small effect on average entry length as well.

The entry length distribution is plotted on Figure 6.4. Despite the absence of one-word-entries, the points are nicely fit by a straight line confirming the nice approximation by a lognormal distribution.

It will be recalled that in the previous study of page location distribution for the manually implemented version of the algorithm on Computerized Library Catalogs there was a significant bend in the Zipf-Mandelbrot straight line due to either a reduced number of singly occurring entries, or an excessive number of multiply occurring entries. For the machine version of the algorithm the page location distribution (or, more accurately, the abstract number location distribution) does not show this deviation (see Figure 6.5). The distribution is given in Table 6.2.

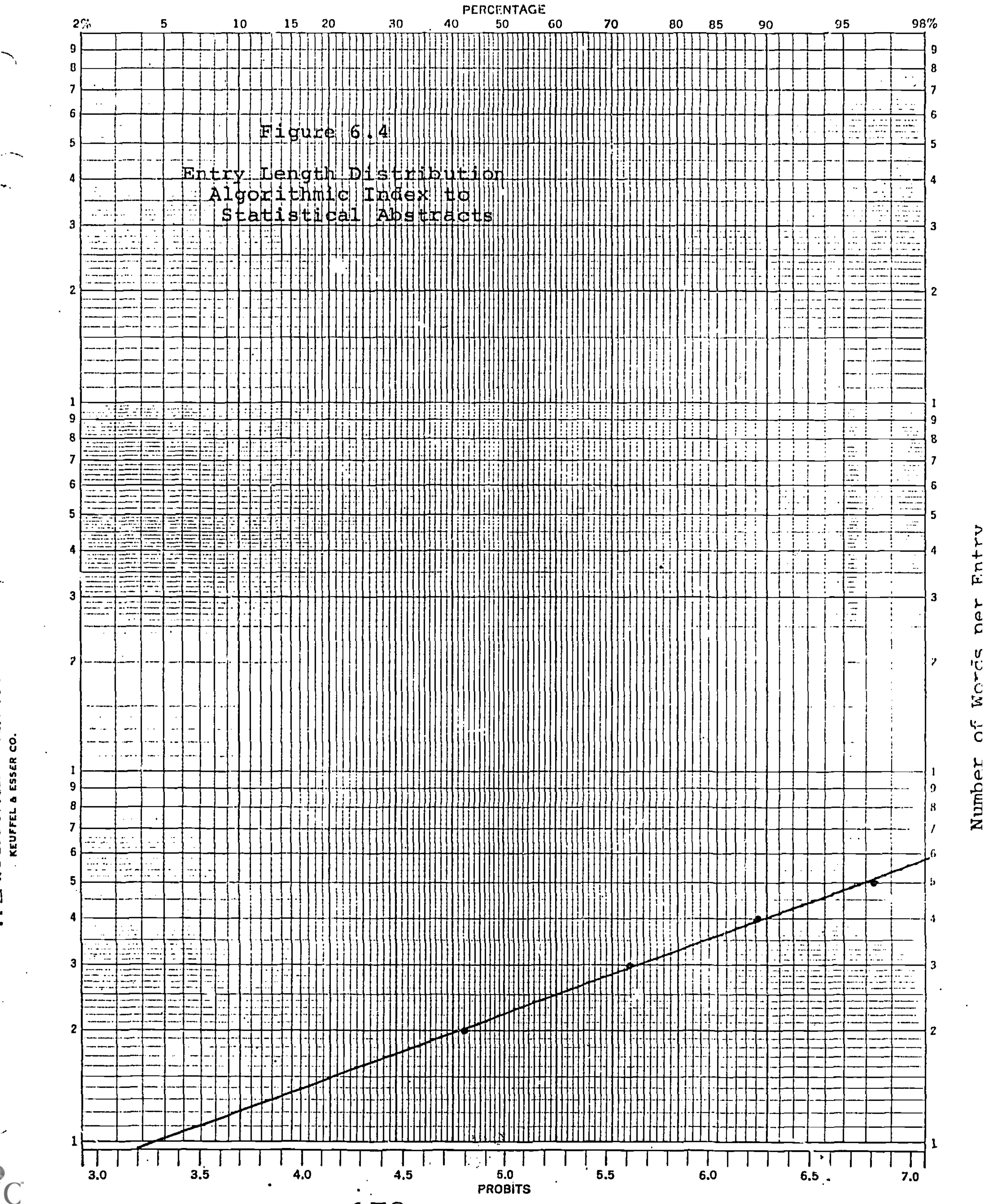


Figure 6.4

Entry Length Distribution
Algorithmic Index to
Statistical Abstracts

Figure 6.5

Abstract Number Location Distribution
Algorithmic Index to Statistical Abstracts

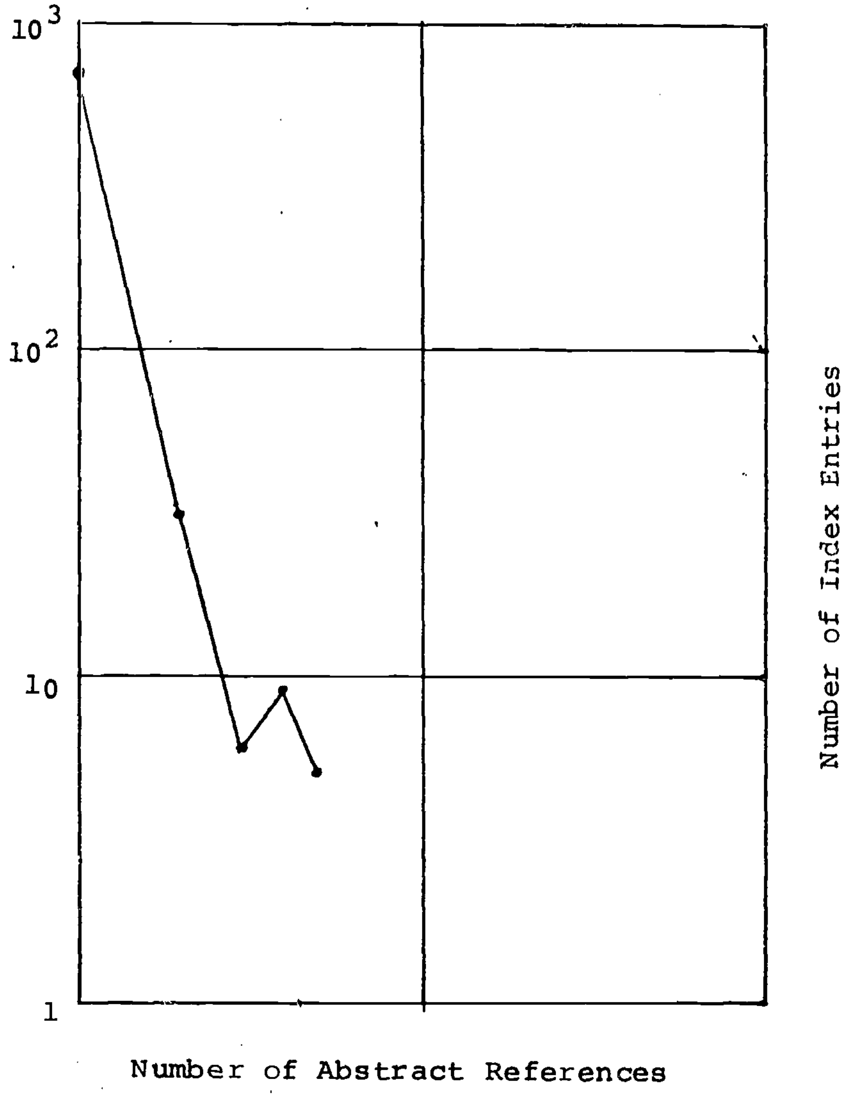


Table 6.2

Abstract Number Location Distribution, Algorithmic Index to Statistical Abstracts

Number of Abstract Locations per Entry	Frequency	Cumulative Frequency
1	703	703
2	31	734
3	6	740
4	9	749
5	5	754

When compared to the comparable data for the manual implementation, it is clear that not only has the difficulty of an insufficient proportion of singly occurring entries been been corrected by the insertion cf structure word logic, but the slope of the line has been significantly increased from 2.17 to 4.49. This increased slope can of course be attributed in part to the nature of the material covered in the two cases and, perhaps, in greater proportion to the structure of the material (i.e. fifty abstracts vs. a single text). Nonetheless, the increase in slope does tend to confirm the expectation that use of structure-word entries is desirable to increase slope.

Potentially more useful than the amalgamation of indexes to abstracts to papers or books is the amalgamation of indexes to the primary texts themselves. We have undertaken an extensive project designed to provide a realistic test of the utility of amalgamations of book indexes as well as an indication of the problems that would be encountered in the preparation of such access mechanisms.

The indexes contained in 80 books on statistics have been committed to machinable form. Approximately 30,000 index entries (not all of which are distinct) are represented, which is nearly 400 entries per book. This is significantly less than the average of 838 index entries per book obtained from the Fondren Index Sample, but, as is clear from Table 4.4, it is well within the deviations typically obtained by restriction of a sample to small and specially defined subsets. We have not attempted to determine the average number of pages per book in this statistics sample; it may well be that the average number of index entries per page is in closer agreement with the figure obtained for the Fondren Index Sample.

The Statistics Index Sample is currently in the early stages of amalgamation. In this report we can only exhibit a combined alphabetically ordered list which has not been formatted (to reproduce the usual format of a book index) and which exhibits the consequences of some program "bugs" not yet corrected which result in the replication of input records at various places throughout the amalgamated index. In spite of these difficulties, the amalgamated list is already a valuable access tool.

Table 6.3 lists the books that constitute the Statistical Index Sample. The code in the leftmost column is the abbreviation for the book used in the amalgamated index. These books were chosen by a professional statistician as representative of the more important information in the statistics field that is available in monograph form. The choice of 80 books rather than a larger number is purely conventional; continuation of this project will increase the data base and permit us to determine how the yield of new index terms varies with increasing size of the sample.

Following the lead of the analysis of the structure of the index to a single book given in Chapters 3 and 4, we see that the rank-frequency distribution Figure 6.6 is just another form of the index reference distribution discussed in those chapters; in the form shown here, the abstract entries appear at the top left part of the graph, and the horizontal portions of the graph correspond to those entries which refer to the same number of text locations. Consequently, the abstract entries for the Statistics Sample certainly include those that have ranks less than 30, and may include several more but not any with rank greater than 50.

Table 6.4 lists the 30 index terms that refer to the greatest number of pages; personal names have been placed in the right hand column; otherwise the order of appearance in the amalgamated index list is the order shown in the table.* This list is a useful pedagogical tool, providing

* The frequencies given here are very tentative, as no attempt has yet been made to agglomerate proper names appearing in variant form.

TABLE 6.3

Bibliographical Description of the Statistics Index Sample

A	Elementary Decision Theory - Chernoff
B	Nonparametric Methods in Statistics - Fraser
C	Statistical Methods for Chemists - W. J. Youden
D	Analysis of Straight-line Data - Acton
E	Testing Statistical Hypotheses - E. L. Lehmann
F	Introduction to Mathematical Statistics - Paul G. Hoel
G	The Design and Analysis of Experiments - O. Kempthorne
H	An Introduction to Multivariate Statistical Analysis - T. W. Anderson
I	Statistics--An Introduction - D. A. S. Fraser
J	Linear Computations - Paul S. Dwyer
K	Modern Probability Theory and Its Applications - Parzen
L	Planning of Experiments - Feller
M	Theory of Games and Statistical Decisions - Blackwell and Girshick
N	An Introduction to Probability Theory and its Applications, Vol 1 - Feller
O	Elementary Statistics - Paul G. Hoel
P	The Elements of Probability - Cramer
Q	Statistical Decision Theory - Weiss
R	Introduction to Probability and Random Variables - Wadsworth and Bryan
S	Introduction to the Theory of Statistics - Mood and Graybill
T	Elements of Probability and Statistics - Wolf
U	An Introduction to Linear Statistical Models, Vol. 1 - Graybill
V	Elements of the Theory of Markov Processes and Their Applications - Bharucha-Reid
W	Geometrical Probability - Kendall and Moran
X	Fundamentals of Statistical Reasoning - Quenouille
Y	Characteristic Functions - Lukas
Z	An Introduction to Probability Theory and Its Applications, Vol. 2 - Feller
AB	Elements of Mathematical Statistics - Alexander
AC	Statistical Theory and Methodology in Science and Engineering - Brownlee
AD	Statistics and Experimental Design, Vol 1 - Johnson and Leone
AE	Mathematical Statistics - Wilkes
AF	Experimental Designs - Cochran and Cox
AI	A Course in Probability Theory - Kai Lai Chung
AJ	Essentials of Probability - Arthur Yaspan
AK	The Design of Experiments - Fisher
AL	Computational Handbook of Statistics - Bruning and Kintz
AM	Design and Analysis of Experiments - Quenouille
AN	Handbook of Statistical Tables - Owen
AO	The Elements of Probability - Berman
AP	Design and Analysis of Industrial Experiments - Davies
AQ	Statistical Theory - Lindgren
AR	Introduction to Statistics - Carlborg
AS	Probability and Statistics - Adler and Rossler
AT	Measuring Uncertainty--An Elementary Introduction to Bayesian Statistics
AU	A Brief Introduction to Probability Theory
AV	Statistical Design and Analysis of Experiments for Development Research - Villars
AW	Statistics in Research - Ostle
AX	Schaums Outline Series Theory and Problems of Probability - Lipschutz
AY	Elementary Mathematical Programming - Metzger
AZ	Statistical Inference for Markov Processes - Billingsley

Continued -

BC Introduction to Probability--A Programmed Unit in Modern Mathematics
BD Statistical Analysis of Stationary Time Series - Grenader and Rosenblatt
BE Statistical Methods in Experimentation--An Introduction - Lacey
BF Stochastic Processes--Basic Theory and Its Application - Prabhu
BG Probability and Frequency - Plummer
BH Statistical Methods for Research Workers
BI Probability, an Intermediate Text-Book - Bixley
BJ Regression Analysis - Williams
BK Statistical Processes and Reliability Engineering - Chorafass
BL Introduction to Probability and Mathematical Statistics - Birnbaum
BM Elementary Mathematical Statistics - Baten
BN Introduction to Biostatistics - Bancroft
BO Sampling Techniques - Cochran
BP A History of the Mathematical Theory of Probability - Todhunter
BQ Statistical Methods in Biology - Bailey
BR Statistical Theory - The Relationship of Probability Credibility and Error
BS An Introduction to Multivariate Statistical Analysis - Anderson
BT Probability and Experimental Errors in Science - Parratt
BU Contributions to Order Statistics - Sarhan and Greenberg
BV Introduction to Statistical Method - Ehrenfeld and Littauer
BW Theory of Probability - Jeffreys
BX Statistical Adjustment of Data - Deming
BY Statistical Analysis in Chemistry and the Chemical Industry - Bennett and Franklin
BZ Probability Random Variables and Stochastic Processes - Papoulis
CD Elements of Queueing Theory with Applications - Saaty
CE Stochastic Processes - Doob
CF Sample Survey Methods and Theory Vol 1 Methods and Applications Hansen, Hurwitz
CG Advanced Statistical Methods in Biometric Research - C Radhakrishna Rao
CH Introduction to the Mathematics of Statistics - Robert W. Burgess

Figure 6.6

Rank - Frequency of Reference Distribution Statistical Index Sample

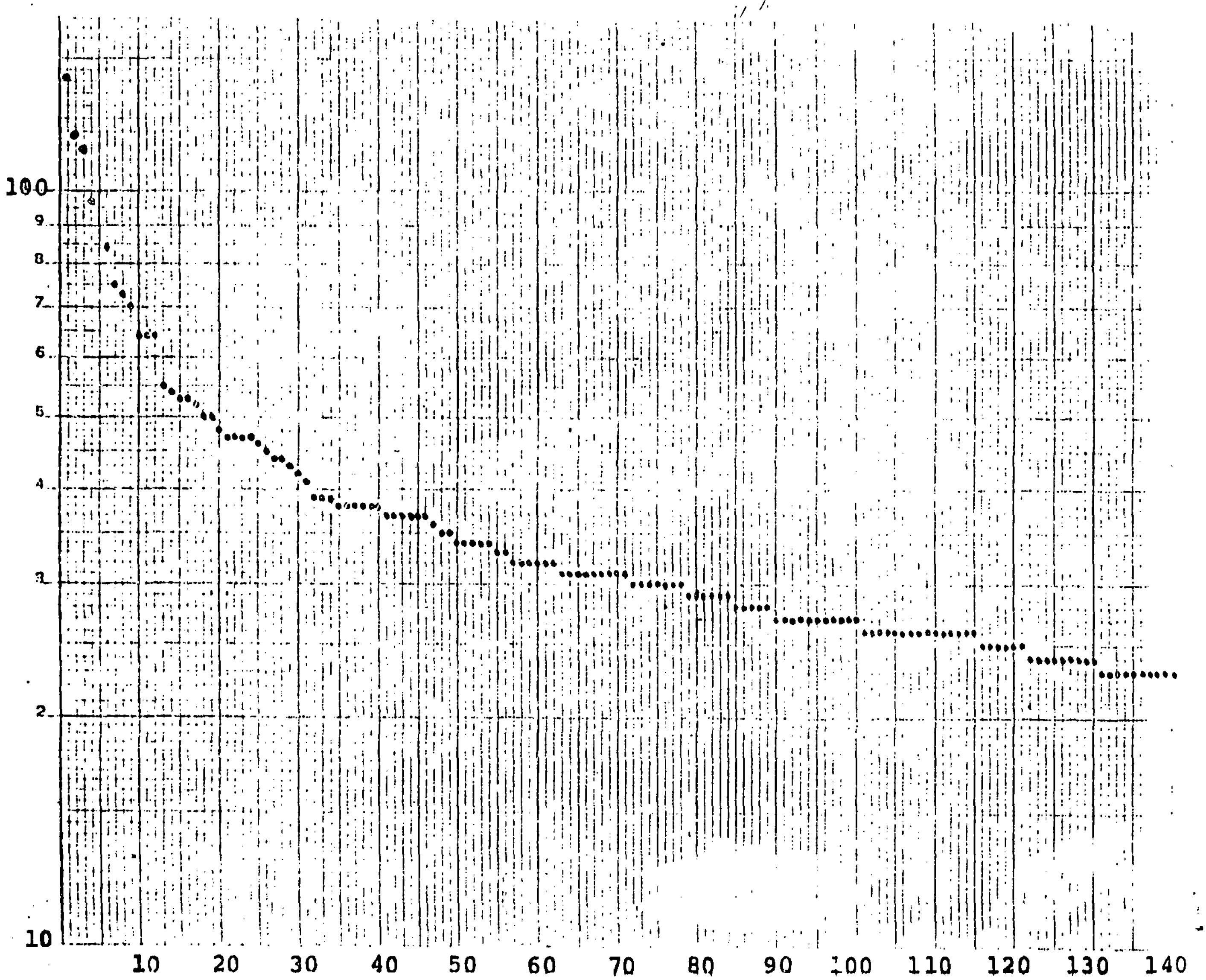


Table 6.4

ABSTRACT ENTRIES FOR THE
AMALGAMATED STATISTICS INDEX SAMPLE

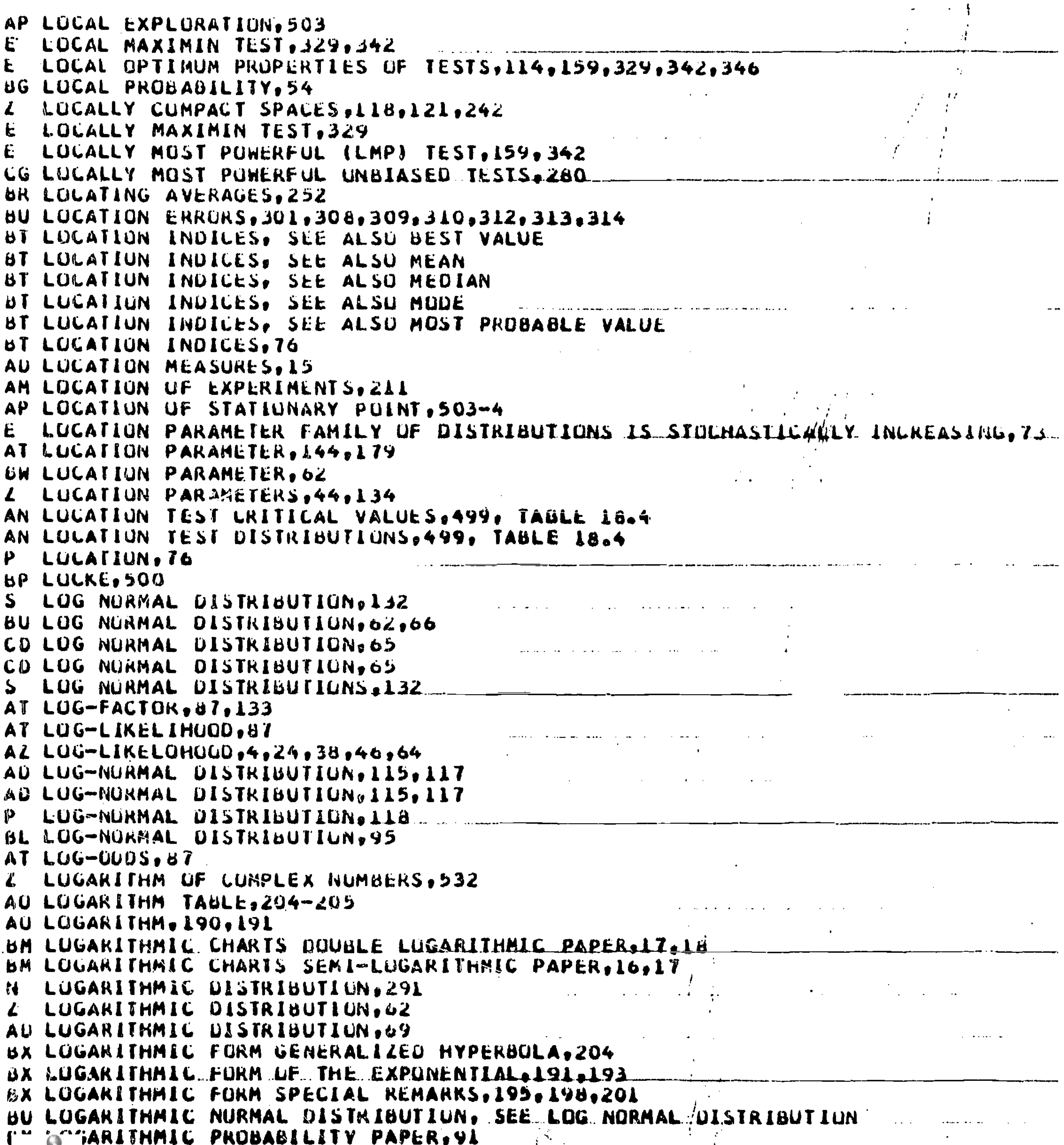
Normal distribution	Fisher, R. A.
Binomial distribution	Student
Poisson distribution	Pearson, E. S.
Degrees of freedom	Kendall, M. G.
Conditional probability	Bartlett, M. S.
Standard deviation	Cramer, H.
Analysis of variance	Neyman, J.
Distribution	
Chi-square distribution	
Central limit theorem	
Least squares	
Variance	
Correlation coefficient	
Median	
Cauchy distribution	
Covariance	
Independence	
Random variable	
Exponential distribution	
Gamma distribution	
Moments	
Bivariate normal distribution	
Multinomial distribution	

as it does an immediate and objective overview of the important subjects in statistics as well as the important contributors. It plays the same role relative to that portion of the field of statistics represented in the monograph literature that the abstract entries for the books described in Chapter 5 played; and it increases the degree of information compression as well.

Figure 6.7 shows one page from the uncorrected form of the amalgamated Statistics Index Sample described above. This page has been selected to include the entry "log normal" and those related to it. Observe that six books (coded P, S, AD, BL, BU, CD) contain references to the log normal distribution; since this represents only 7.5% of the books in the Statistics Index Sample, the unsophisticated inquirer will realize a very significant saving in search time with a reasonable degree of assurance that most of the significant references will either be covered directly within these six books, or more comprehensive treatments will be noted in their bibliographies.

Figure 6.7

Sample Page from the Amalgamated Statistics Index



```
AP LOCAL EXPLORATION,503
E  LOCAL MAXIMIN TEST,329,342
E  LOCAL OPTIMUM PROPERTIES OF TESTS,114,159,329,342,346
BG LOCAL PROBABILITY,54
Z  LOCALLY COMPACT SPACES,118,121,242
E  LOCALLY MAXIMIN TEST,329
E  LOCALLY MOST POWERFUL (LMP) TEST,159,342
CG LOCALLY MOST POWERFUL UNBIASED TESTS,280
BR LOCATING AVERAGES,252
BU LOCATION ERRORS,301,308,309,310,312,313,314
BT LOCATION INDICES, SEE ALSO BEST VALUE
BT LOCATION INDICES, SEE ALSO MEAN
BT LOCATION INDICES, SEE ALSO MEDIAN
BT LOCATION INDICES, SEE ALSO MODE
BT LOCATION INDICES, SEE ALSO MOST PROBABLE VALUE
BT LOCATION INDICES,76
AU LOCATION MEASURES,15
AM LOCATION OF EXPERIMENTS,211
AP LOCATION OF STATIONARY POINT,503-4
E  LOCATION PARAMETER FAMILY OF DISTRIBUTIONS IS STOCHASTICALLY INCREASING,73
AT LOCATION PARAMETER,144,179
BW LOCATION PARAMETER,62
Z  LOCATION PARAMETERS,44,134
AN LOCATION TEST CRITICAL VALUES,499, TABLE 18.4
AN LOCATION TEST DISTRIBUTIONS,499, TABLE 18.4
P  LOCATION,76
BP LOCKE,500
S  LOG NORMAL DISTRIBUTION,132
BU LOG NORMAL DISTRIBUTION,62,66
CD LOG NORMAL DISTRIBUTION,65
CD LOG NORMAL DISTRIBUTION,65
S  LOG NORMAL DISTRIBUTIONS,132
AT LOG-FACTOR,87,133
AT LOG-LIKELIHOOD,87
AZ LOG-LIKELOHOOD,4,24,38,46,64
AD LOG-NORMAL DISTRIBUTION,115,117
AD LOG-NORMAL DISTRIBUTION,115,117
P  LOG-NORMAL DISTRIBUTION,118
BL LOG-NORMAL DISTRIBUTION,95
AT LOG-ODDS,87
Z  LOGARITHM OF COMPLEX NUMBERS,532
AU LOGARITHM TABLE,204-205
AU LOGARITHM,190,191
BM LOGARITHMIC CHARTS DOUBLE LOGARITHMIC PAPER,17,18
BM LOGARITHMIC CHARTS SEMI-LOGARITHMIC PAPER,16,17
N  LOGARITHMIC DISTRIBUTION,291
Z  LOGARITHMIC DISTRIBUTION,62
AU LOGARITHMIC DISTRIBUTION,69
BX LOGARITHMIC FORM GENERALIZED HYPERBOLA,204
BX LOGARITHMIC FORM OF THE EXPONENTIAL,191,193
BX LOGARITHMIC FORM SPECIAL REMARKS,195,198,201
BU LOGARITHMIC NORMAL DISTRIBUTION, SEE LOG NORMAL DISTRIBUTION
C  LOGARITHMIC PROBABILITY PAPER,91
```

Appendix I

Abstract Index Entries:
A Uniform Sample from the
Fondren Index Sample

25
BF181.M3 1931

Marston, William Moulton
Integrative Psychology

Sum= 1432 / 29.54 ≐ 48

23 Marston, W.M.

17 Freud, Sigmund

15 Watson, J.B.

13 Cannon,. W.B.

11 Adler, Alfred

10 Desire
10 Jung, Carl
10 Libido
10 Woodsworth, R.S.

9 Compliance
9 Passion

8 Allport, F.H.
3 Behaviourism
8 James, WM.
8 MacDougall, Wm

7 Herrick, C.J.
7 Satisfaction
7 Sherrington, C.S.

6 Captivation
6 Dominance
6 Psychoanalysts

5 Carlson, A.J.
5 Erotic drive
5 Inducement
5 Passion response
5 Visual discrimination, substances, hypothetical

4 Angell, J.R.
4 Archtypes
4 Cell body
4 Compliance, motives
4 Eng, H.
4 Hering, E.
4 James-Lange, Theory of Motion
4 Law of integrative sequence
4 Origination response
4 Passion motives
4 Submission
4 Trolaut, L.T.
4 Unit responses, compound
4 Washburn, M.F.
4 Yerkes, R.M.

50

McLean, Archibald BV2532.M3 1921

The History of the Foreign Missionary Society

Sum= 498 / 29.54 = 16.86

7	Fallen, The
6	Moore, W.T.
4	Nurses being trained
3	Bilaspur
3	Johnson, Miss Kate V.
3	Loos, C.L.
3	Moore, W.T., Quoted
3	Rijnhardt, Dr. Susie C.

75
Coolidge, Archibald Cary D443.C6 1927
Ten Years of War and Peace

Sum= 415 / 29.54 = 14

31 Great Britain mentioned

20 France, mentioned
20 Poland

19 League of Nations, mentioned

18 Versailles, Treaty of

16 Wilson, Woodrow

15 Hungary

13 Algeria
13 Hughes, Charles E., Secretary of State

11 Germany, mentioned
11 Harding, Warren G.
11 Morocco

10 China
10 France, estrangement between and Great Britain
10 Japan,mentioned
10 Rumania

100
Sackville-West, Victoria Mary DA690.K7 1922
Knole and the Sackvilles

Sum= 307 / 29.54 = 10

7 Sackville, Lady Margaret(afterwards Countess of Thanet), mentioned in Lady Anne Clifford's diary

4 Pepys, Samuel, quoted
4 Walpole, Horace, quoted on Knole

3 Devonshire, Duchess of, his (i.e. 3rd Duke of Dorset) letter to her
3 Dryden, John, his debt to 6th Earl of Dorset
3 Gorboduc
3 Macaulay, quoted
3 Sackville, Charles, 6th Earl of Dorset, songs quoted
3 Sackville, Lord George, quoted
3 Wraxall, Sir Nathaniel, quoted

125
DF521.S4 1966

Sherrard, Philip
Byzantium

Sum= 1643 / 29.54^2 = 1.88

10 Churches: in Constantinople
10 Frescoes

150

Institutc of Culture
The Cultural Heritage of India

Sum= 4906 / 29.54^2 = 5.6

51	Kṛṣṇa (śrī)
46	Siva
41	"Bhagavad- Gītā"
37	Visnu
33	Brahman
32	Guru(s)

175
E178.6.S3 1965

Saveth, Edward, ed.
Understanding the American Past

1958 / $29.54^2 = 2.24$

28 Beard, Charles A. and Mary

20 Jefferson, Thomas

17 Turner, Frederick Jackson

200

Link, Arthur S. E741.L55 1963

American Epoch: A History of the United States Since the 1890's

Sum= 7016 / 29.54^2 = 8

14 Prices: agricultural

12 Foreign relations: Anglo-American

11 Federal income tax: individual
11 Tax: individual income

10 Farmers, income of
10 Legislation:agricultural

9 Agriculture, legislation for
9 Railroads: rates of

225
F864.b68 v4 1930

Bolton, Herbert Eugene
Anza's California Expedition

Sum= 648 / 29.54 = 22

132	Mass
111	Anza, Juan Bautista
60	Garces, Fray Francisco
46	Monterrey
44	San Gabriel Mission
27	San Diego mentioned.
26	Colorado River
26	Ribera (Rivera) Fernando de.
26	Sierra Madre de California
23	Apaches
21	Palma, Salvador
21	Sierra Nevada
20	San Miguel de Horcasitas
19	Eixarch (Eyxarch) Fray Thomàs
18	San Francisco, harbor and settlement
17	Mexico
16	Spaniards
15	Christian Indians
15	Fages, Pedro
15	Gila River
14	Pablo (Captain Feo) Yuma Chief
13	Crespi, Fray Juan
13	Rio de San Francisco (San Joachin)

250
De Garmo, Ernest Paul HB199.W595 1960
Engineering Economy

Sum= 650 / 29.54 = 22

5 Terborgh, Genge

4 Break even charts, examples of

3 Balance sheet, example of
3 Deferred-investment studies, examples of
3 Minimum cost point
3 Personnel factors, lighting
3 Rate of return, determination of
3 Selection, of design
3 Survivor curves, examples of

2 Accidents, effect of lighting on
2 Annuities whose present value is
2 Borrowed capital, cost of
2 Bureau of Internal Revenue relation to depreciation
2 Capital gains, and losses
2 Capital gains and losses, carry-over of
2 Capitalized cost, example of application
2 Costs, accuracy of estimates of
2 Costs, labor
2 Depreciation, sum-of-the-years'-digits
2 Hoover Dam
2 Income and expense statements, example of
2 Income taxes in public utility studies
2 Increment costs
2 Labor, turnover of
2 Life, economic
2 Life, useful
2 MAPI replacement formulas, forms for use in
2 Material, selection of
2 Multiple-purpose works, evaluation of benefits from
2 Overhead expense bases for distribution of
2 Plant location, economy studies of
2 Power factor, effect on utility rates
2 Rate schedules, block demand
2 Rautenstrauch, Walter
2 Risk, factors affecting
2 Selection of methods or processes
2 Self liquidating projects, relation of taxes to
2 Self liquidating projects, repayment of capital in
2 Wage payment, piece work

275
HD20.C554 1958

Chorafas, Dimitris N.
Operations Research for Industrial Management

Sum= 141 / 29.54 = 5

17 Charts on simulated business results

11 Computers usage
11 Simulation

10 Allocation

9 Managerial decisions

300
HF2046.S62 1904

Smart, William

The Return to Protection

S= 201 / 29.54 = 7

20	Chamberlain, J.
20	Germany
19	Board of Trade
14	France
10	Giffen, Sir Robert
10	Shipping
9	America
9	America and protection
9	Canada

325
HM66.C7 1920

Cole, George Howard Douglas
Social Theory

Sum= 382 / 29.54 = 13

22	Trade Unions
15	"State, The"
14	Associations
12	Churches
12	Functional Equity, Court of, organisation
10	Law
10	Rousseau
9	Sovereignty
8	Function in relation to individual, perversion of
8	Marxism
8	Will, as a basis of Society
7	Middle Ages
7	Parliament

350
JA84.R9 U8 1964

Utechin, Sergei
Russian Political Thought

Sum= 409 / 29.54 = 14

50	Economy
45	Classes, social
43	Law
39	Germany
34	Individualism
34	Monarchy
33	Emigration
33	France
33	Peasantry
32	Intelligentsia
31	Education
30	Property
30	Terror
29	Christianity
29	Culture
29	Equality
29	Moscow
29	Nationalism
29	Nobility
29	St. Petersburg

Cooper, Lane
Two Views of Education

Sum= 775 / 29.54 = 26

43	America
42	Milton
39	Plato
34	Shakespeare
27	Aristotle
26	Homer
26	Teacher (of English, etc.)
25	Greek, Study of
24	Middle Ages
22	Horace
22	Wordsworth
21	Bible
20	Latin, Study of
18	Cicero
17	Rewards of the Teacher
16	Odyssey
16	Rome
16	Virgil
15	Chaucer
15	Discipline
15	England
15	Socrates
14	Dante
13	Democracy
13	Greece
13	Rousseau

400
Morgan, Alexander
LC191.M6 1916
Education and Social Progress

Sum= 254 / 29.54 = 9

8 Children, diseases of
8 Education, practical
8 Inter-Departmental committee

7 Kindergartens
7 Practical education
7 Vocational education

6 Commission, Royal, on Poor Laws
6 Continuation education
6 Edinburgh, continuation schools
6 Education, continuation
6 Education, vocational
6 Education and health
6 Plato
6 Scotch Education Deptartment
6 Slums, children in

425

Tomkins, Calvin ND553.D774 T6 1965

The Bride and the Bachelors

Sum= 414 / 29.54 = 14

22	"Bride Stripped Bare By Her Bachelors, Even, The" (Duchamp)
15	Tudor, David
14	Cunningham, Merce
13	Rauschenberg, Robert
12	Klüver, Billy
11	Johns, Jasper
10	Duchamp, Marcel
8	Feldman, Morton
8	Thomson, Virgil
8	Tinguely, Jean
7	Arensberg,Walter C.
7	Breton, Andre
7	Cage, John
7	Cage, Mrs. John
7	Cowell, Henry
7	Dreier, Katherine
7	Kashevaroff, Xenia Andreevna
7	"Nu Descendant un Escalier" (Duchamp)
7	"Nude Descinding a Staircase" (Duchamp)
7	Schönberg, Arnold

450
PN 2598.k4 b6 1931

Bobbe, Dorothie (De Bear)
Fanny Kemble

Sum= 384 / 29.54 = 13

42	Butler, Pierce
40	St. Leger, Harriet
36	Kemble, Charles
22	Butler, Sarah
19	Butler, Fanny
19	Siddons, Sarah
18	Covent Gargen Theatre
18	Lenox,Mass.
17	Kemble, Adelaide
17	Kemble, Mrs Charles
17	Sartoris, Mrs Edward
17	Slavery
15	Kemble, John Mitchell

475
PR2831.H6 1948

Hoppe, Harry Reno
The Bad Quarto of Romeo and Juliet

Sum= 529 / 29.54 = 18

38	Greg, Walter W.
22	Chambers, (Sir) E.K.
18	Hart, Alfred
18	Mc Kerrow, R.B.
14	Burby, Cuthbert
13	Boswell, Eleanor
13	Greene, Robert "Orlando Furioso"
12	Arber, Edward
12	Recollections
12	Shakespeare, William,"The Merry Wives of Windsor"
11	"Orlando Furioso"
10	Anticipations
10	Chamberlain"s Company
10	Danter, John
10	Shakespeare, William "3 Henry VI"
9	"3 Henry VI"
9	"Merry Wives of Windsor, The "
9	Peele, George "The Old Wives" Tale"
9	Peele, George "Edward I"
9	Repetitions

500
Ryals, Clyde, de L PR5588.R9 1964
Theme and Symbol in Tennyson's Poems to 1850

Sum= 322 / 29.54 = 11

37	Keats, John
23	"In Memoriam"
22	William Wordsworth
20	"Two Voices, The"
17	"Palace of Art, The"
15	"Lotus Eaters, The"
15	"Ulyssus"
13	"Mariana"
13	"Recollections of the Arabian Nights"
12	Hallam, Arthur Henry

525
PT7244.K57 1946-49

Kock, Ernst Albin, ed.
Den Norsk-Islandska Skaldediktningen

Sum= 626 / 29.54 = 21

2 Bjark: Bjarkamál, anon.
2 Danir. Danir, anon.
2 Finng: Finngálkn, anon.
2 Jómsvikingar anon.
2 Karlevi: Karlevistenens drottkvädade vers, anon.
2 Oddm.: Oddmjor anon.
2 Rauðsk.: Rauðskeggr. anon.
2 Sveinn tjuguskegg, anon.
2 Svtjúg: Sveinn tjuguskegg, anon.
2 Tångbrand och Gudlev, dikt om, anon.
2 Vagn: Vagn Akason anon.
2 AEvidrápa (orvar-odds): ur Qrvar-odds saga

550

Ostrowski, Alexander QA303.088 v3 1945-54

Vorlesungen Über Differential und Integralrechnung

Sum= 868 / 29.54 = 29

14	Cauchy, A.
9	Euler
6	Cantor, G.
6	Dirichlet
6	Gauss
6	Hardy, G.H.
6	Weierstrass
5	Abel
5	Ellipse
5	Hermite
5	Konvergenzkriterien für uneigentliche Integrale
5	Schwarz, H.A.
4	Bertand, J.
4	Bolzano
4	Cesåro
4	Hausdorff
4	Jensen, J.L.W.V.
4	Newton
4	Pringsheim, A.
4	Riemann, B.
3	Caratheódory
3	Cauchy-Bolzanosches Konvergenzkriterium
3	Chaundy
3	Enveloppe
3	Fresnelsche Integrale
3	Hadamard
3	Inhalt
3	Konvergenzkriterium für unendliche Cauchy -Bolzanoshes
3	Poisson
3	Stieltjes
3	Vergluchskriterium für unendliche -uneigentliche Integrale
3	Zusammenhängend

575
QD9.S71 Ref. 1938

Soule, Byron Avery
Library Guide for the Chemist

Sum= 1833 / 29.54 = 28

5 Gregory

4 Böttger, Wm.
4 Furman, N.H.
4 Water, analysis of

3 Biography, German
3 Browne, C.A.
3 Classen
3 Daniels, F.
3 Dyes, patents on
3 Ferro-alloys, analysis
3 Findlay, A.
3 Glasstone, S.
3 Hahn, D.
3 Hall, W.T.
3 Houben, J.
3 Indexes, patent
3 Koltoff, I.M.
3 Martin, G.
3 Meyer, R.J.
3 Mnemonics
3 Nomenclature, organic
3 Organo-metallic Compounds
3 Ostwald, Wm.
3 Patents, dye
3 Rossman, J.
3 Steel, analysis of
3 Sugar, analysis
3 Thorpe, Edw.
3 Weiser, H.B.
3 Worden, E.C.

Varley, Ernest Reginald
Sillimanite

600
QE391.S5 V3 1965

Sum= 729 / 29.54 = 25

11	Reserves, India
10	Andalusite: U.S.A.
8	Kenya
8	Reserves, U.S.A.
7	Assam, India
7	United States
6	Florida, U.S.A.
6	Georgia, U.S.A.
5	Beneficiation, U.S.A.
5	Bihar, India
5	Brazil
5	California, U.S.A.
5	Dumortierite, U.S.A.
5	Mysore, India
5	Nyasaland
5	South Africa, Republic of
5	Topaz: U.S.A.
4	Aluminium industry
4	Andalusite: U.S.S.R.
4	Andhra Pradesh, India
4	Baker Mountain, Virginia
4	Density
4	Graves Mountain, Georgia
4	Henry Knob, S. Carolina
4	India
4	Kerala, India
4	Kyanite density
4	Lapsa Buru, Bihar
4	Madhya Pradesh, India
4	Maharashtra, India
4	Nevada, U.S.A.
4	New South Wales, Australia
4	Orissa, India
4	Reserves, U.S.S.R.
4	Sillimanite minerals:density
4	South Carolina U.S.A.
4	Transvaal, South Africa
4	United States, National Stockpile Purchase Specification

625

Davis, David Edward
Principles in Mammalogy

Sum= 836 / 29.54 = 28

11	Carnivores
10	Bat(s)
10	Insectivores
10	Woodchucks
9	Marsupials
9	Mutation
9	Teeth
8	Monotremes
8	Whale(s)
7	Herbivores
7	Opossums
7	European rabbit(s)
7	Raccoons
6	Dispersal
6	Maintenance
6	Predators
6	Primates
6	Shrew(s)
6	Vole, meadow
5	Body size, temperature
5	Camels
5	Competition
5	Corpora lutea
5	Elephants
5	Feedback
5	Food
5	Fossils
5	Migration
5	Mole
5	Muskrats
5	Nearctic region
5	Omnivores
5	Oriental region
5	Sex ratio
5	Squirrel(s), ground
5	Temperature

650
RM721.E8 1947

Ewerhardt, Frank Henry
Therapeutic Exercise

Sum= 338 / 29.54 = 11

8 Muscle contraction
8 Paralysis
8 Posture

7 Spastic paralysis, exercise in

6 Flat foot
6 Muscle function volitional tests
6 Poliomyelitis treatment
6 Re-education

5 Hemiplegia
5 Lordosis
5 Poliomyelitis testing by topographical observations
5 Poliomyelitis treatment during acute stage
5 Re-education of upper extremity
5 Scoliosis
5 Upper extremity re-education
5 Zero position

675
Underhill, Charles Reginald TK153.U5 1933
Electrons at Work

Sum= 3827 / 29.54^2 = 4

9 Tube, gaseous-discharge

7 Hertz
7 Light, ultra-violet
7 Maxwell
7 Reaction, Reactors
7 Valence electrons

700
Z5782.A258 Ref. 1954

Bibliography of Medieval Drama
Stratman, Carl Joseph

Sum= 2797 / 29.54^2 = 3.2

44 Passion

37 Comedy

32 Latin

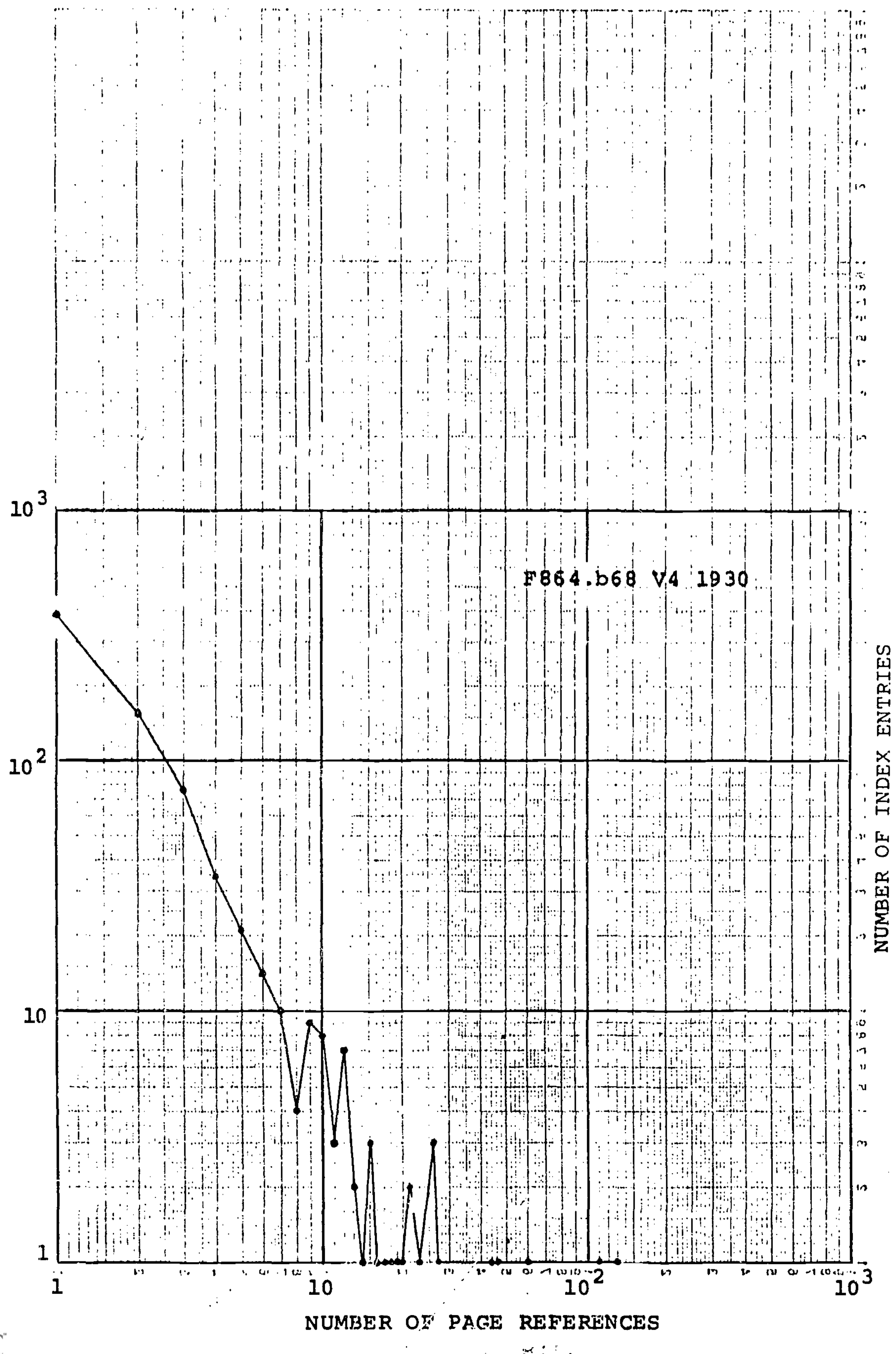
31 Staging

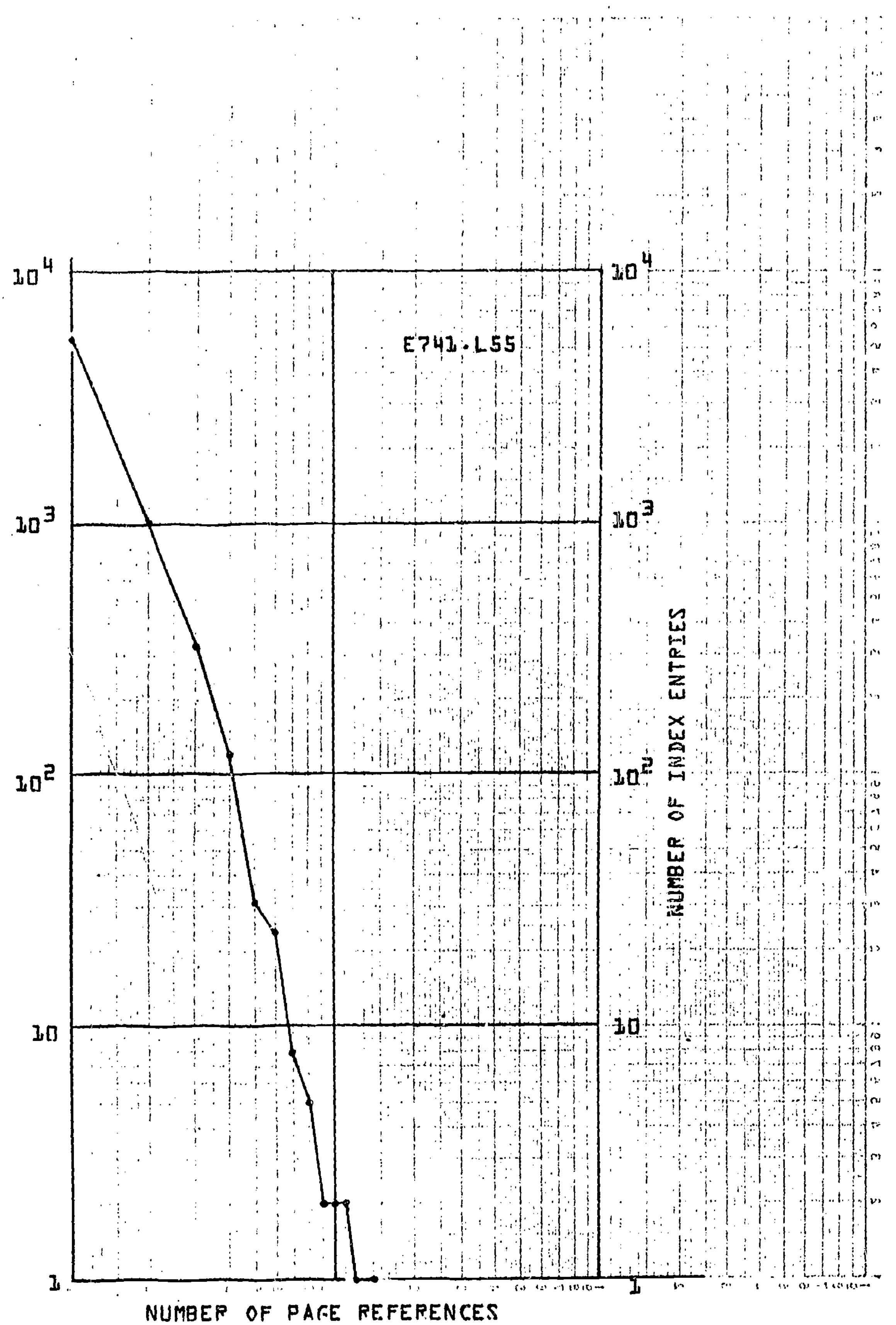
APPENDIX II

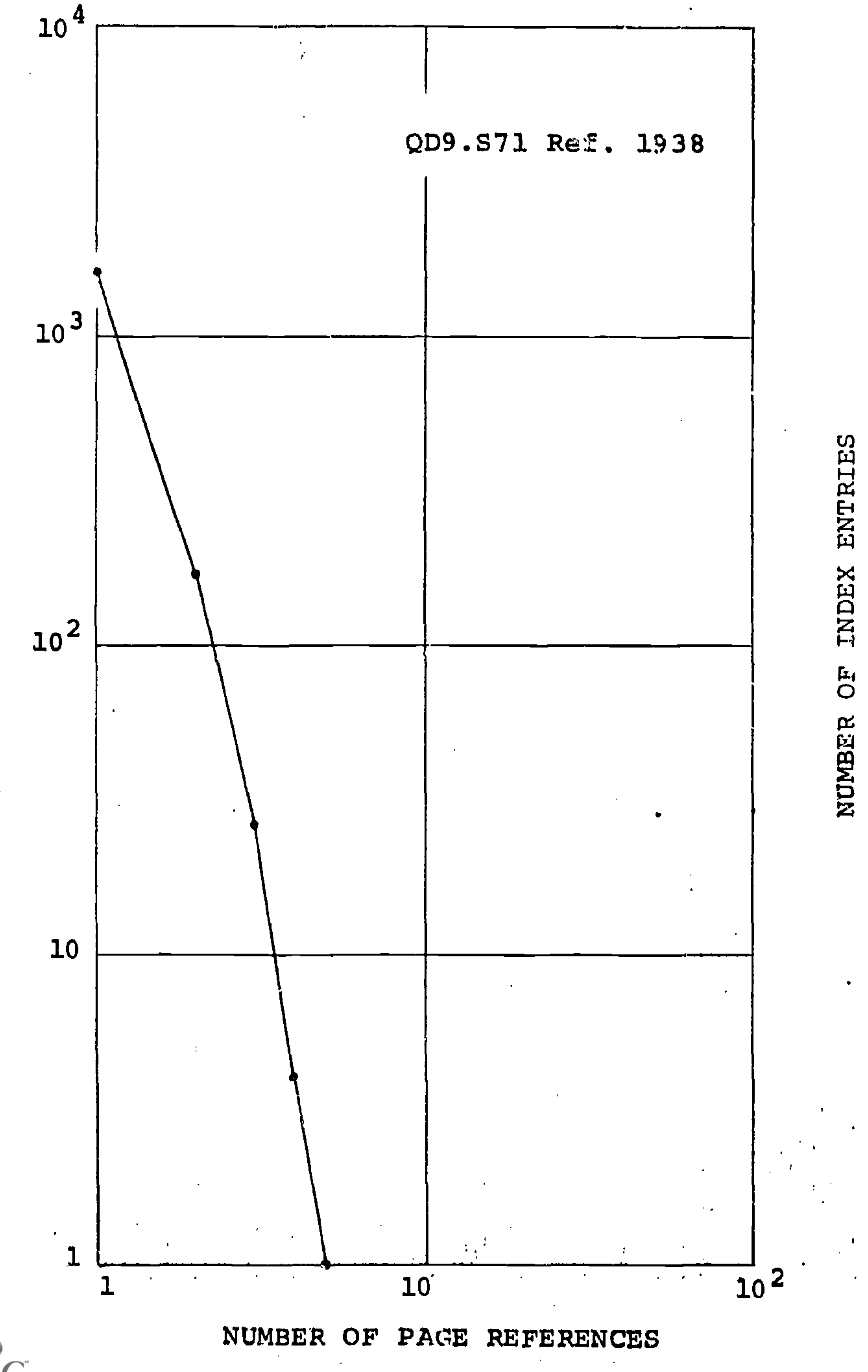
PAGE REFERENCE DISTRIBUTIONS

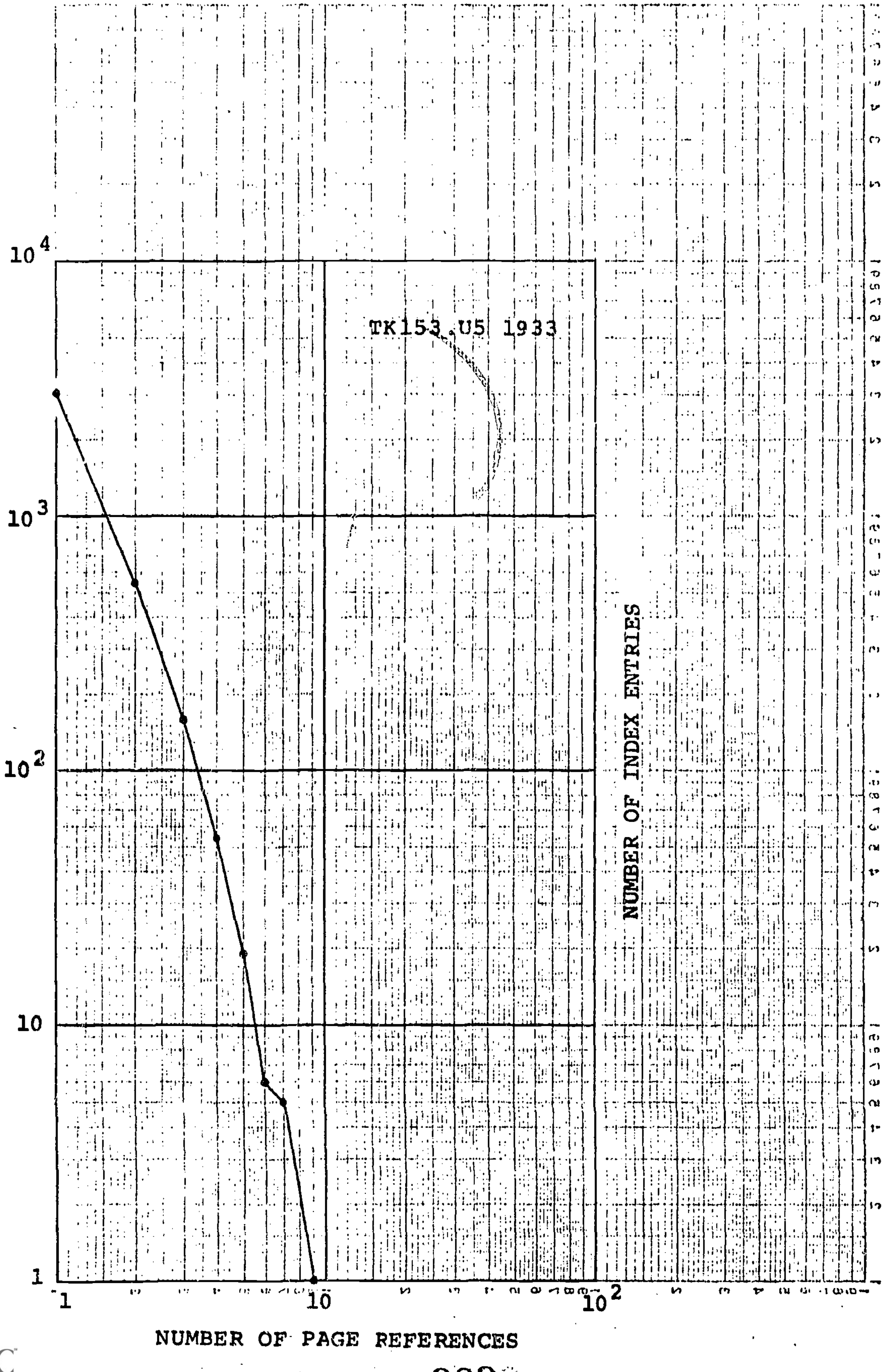
FROM

THE FONDREN INDEX SAMPLE

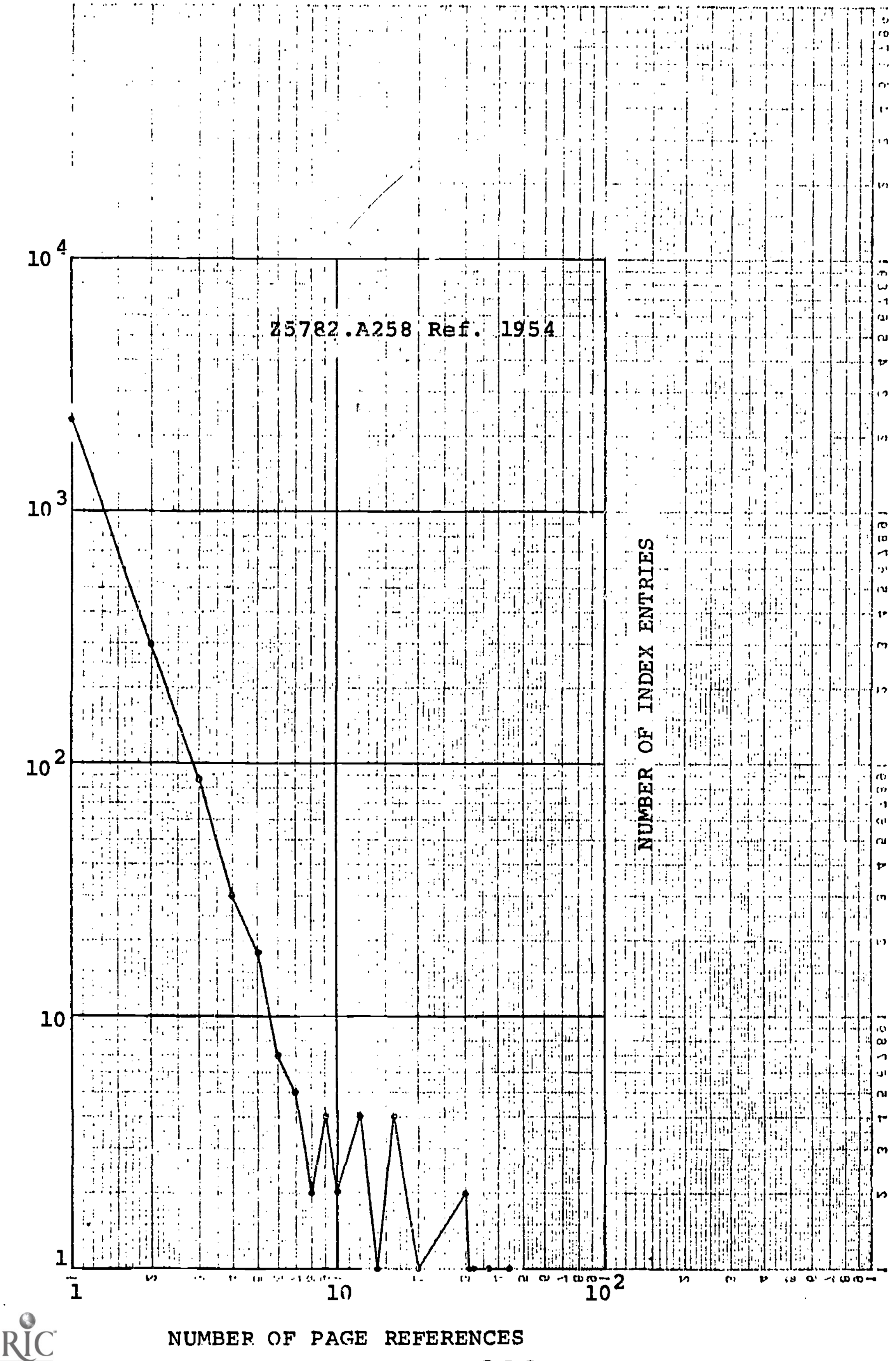

10^3
10^2
10
1
F864.b68 V4 1930
NUMBER OF INDEX ENTRIES
1
10
10^2
10^3
NUMBER OF PAGE REFERENCES

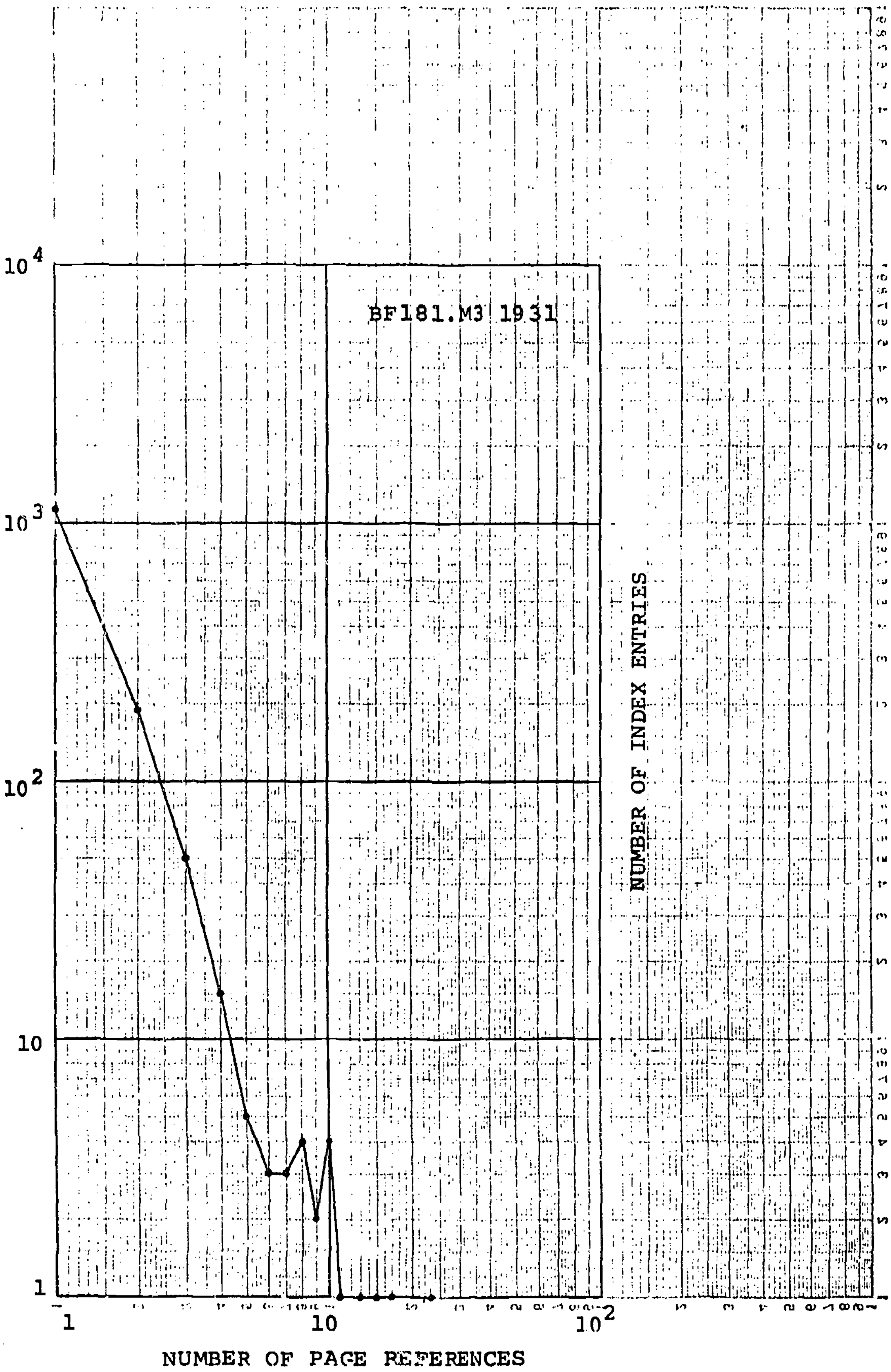

E741.L55
10^4
10^3
10^2
10
1
NUMBER OF INDEX ENTRIES
NUMBER OF PAGE REFERENCES

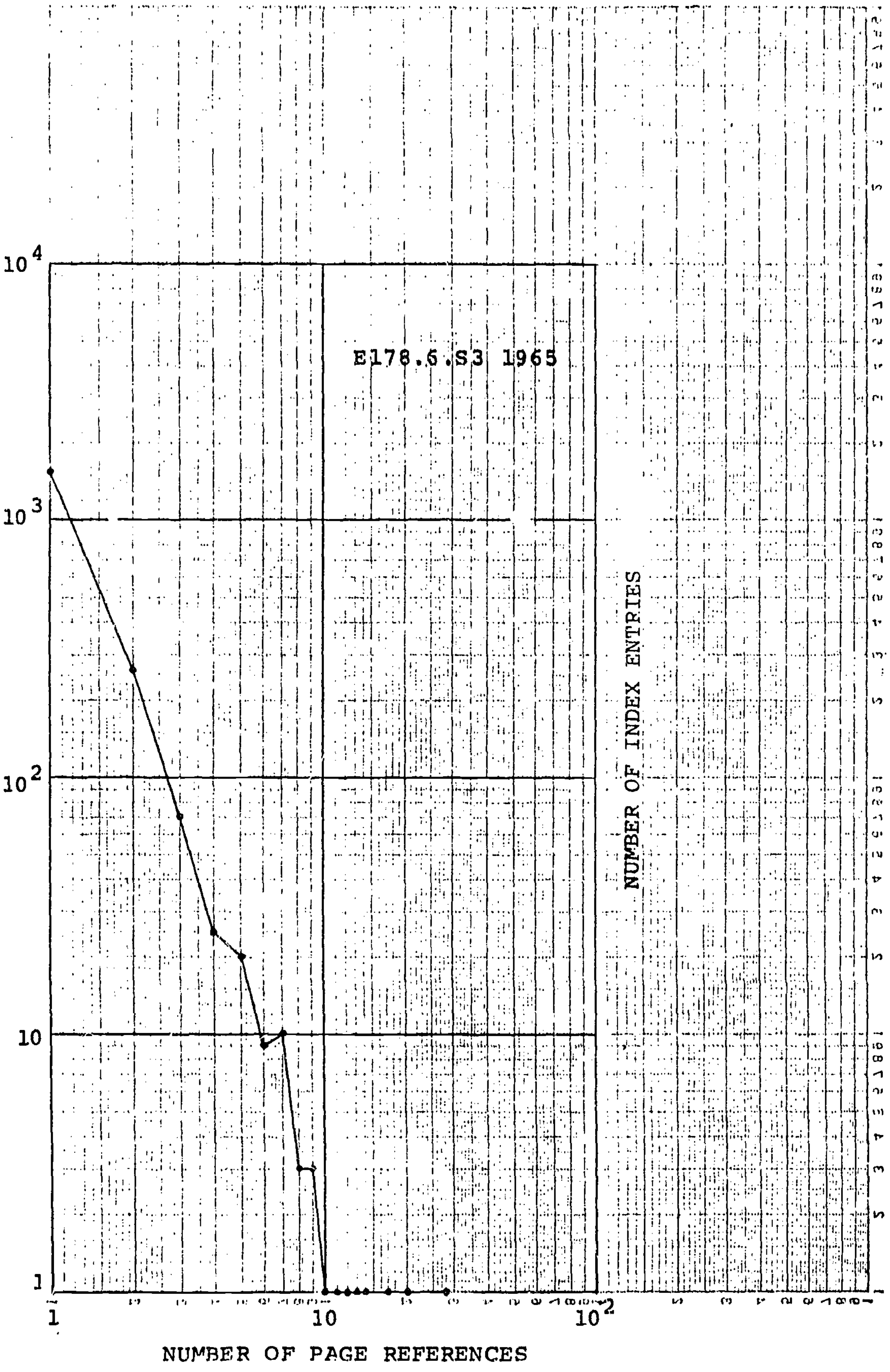

QD9.S71 Ref. 1938
10^4
10^3
10^2
10
1
1
10
10^2
NUMBER OF PAGE REFERENCES
NUMBER OF INDEX ENTRIES

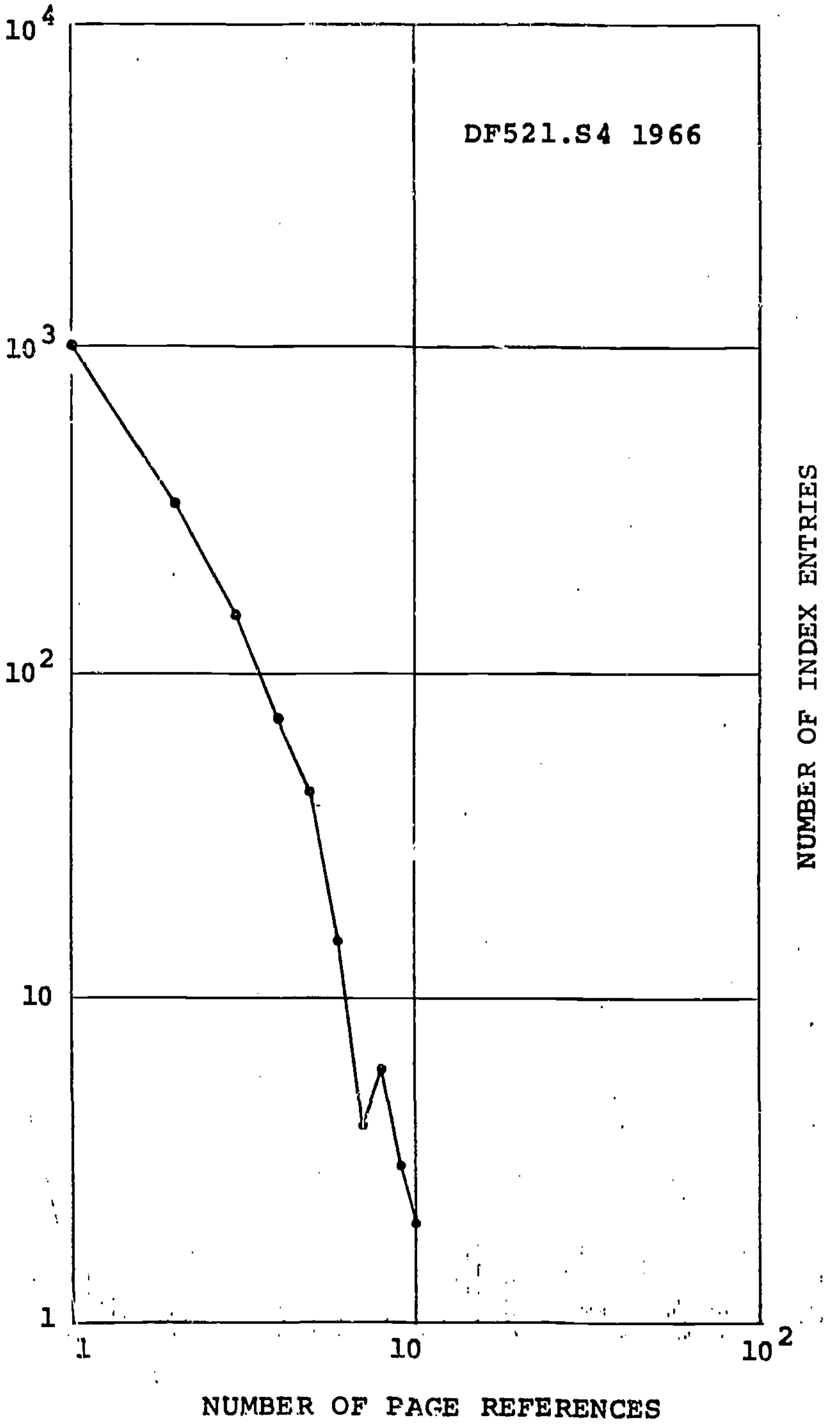


TK153.U5 1933
10^4
10^3
10^2
10
1
1
10
10^2
NUMBER OF INDEX ENTRIES
NUMBER OF PAGE REFERENCES

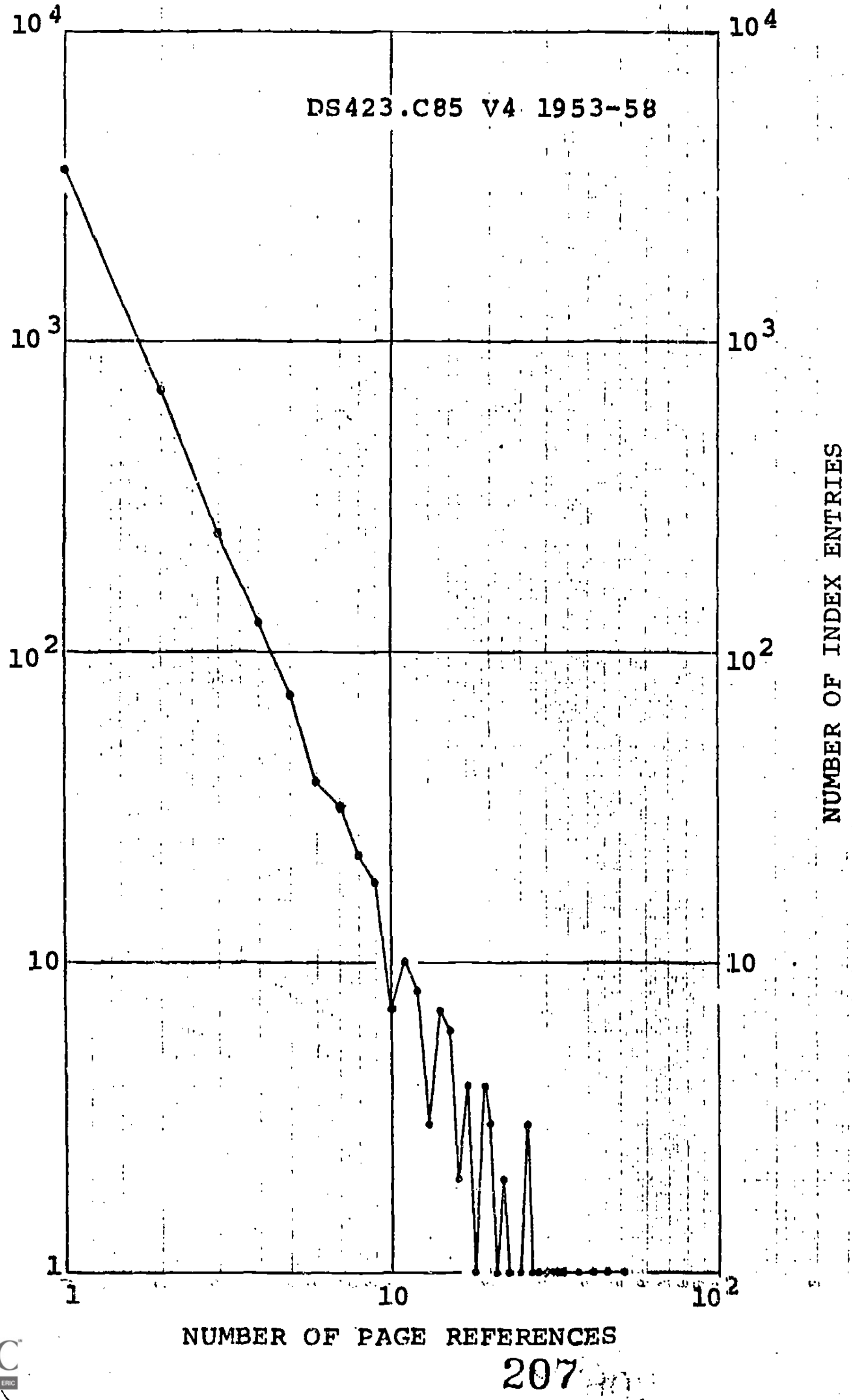

Z5782.A258 Ref. 1954
10^4
10^3
10^2
10
1
NUMBER OF INDEX ENTRIES
1
10
10^2
NUMBER OF PAGE REFERENCES

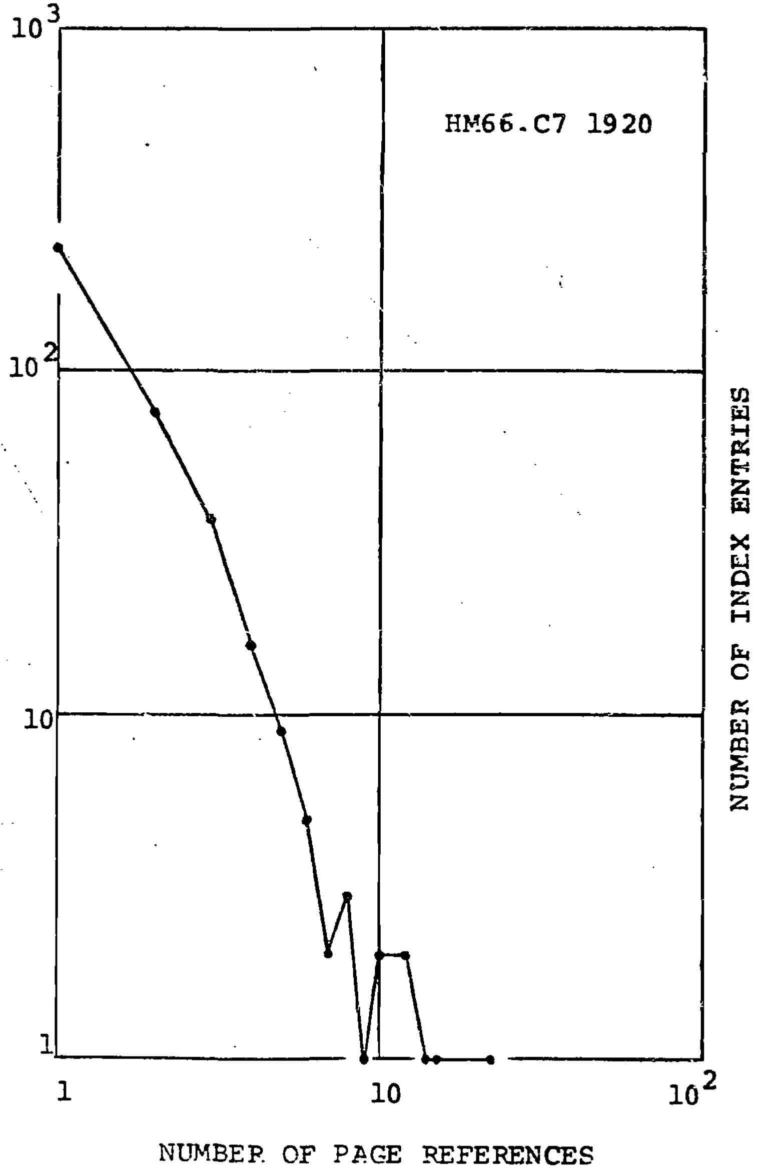

BF181.M3 1931
10^4
10^3
10^2
10
1
1
10
10^2
NUMBER OF INDEX ENTRIES
NUMBER OF PAGE REFERENCES

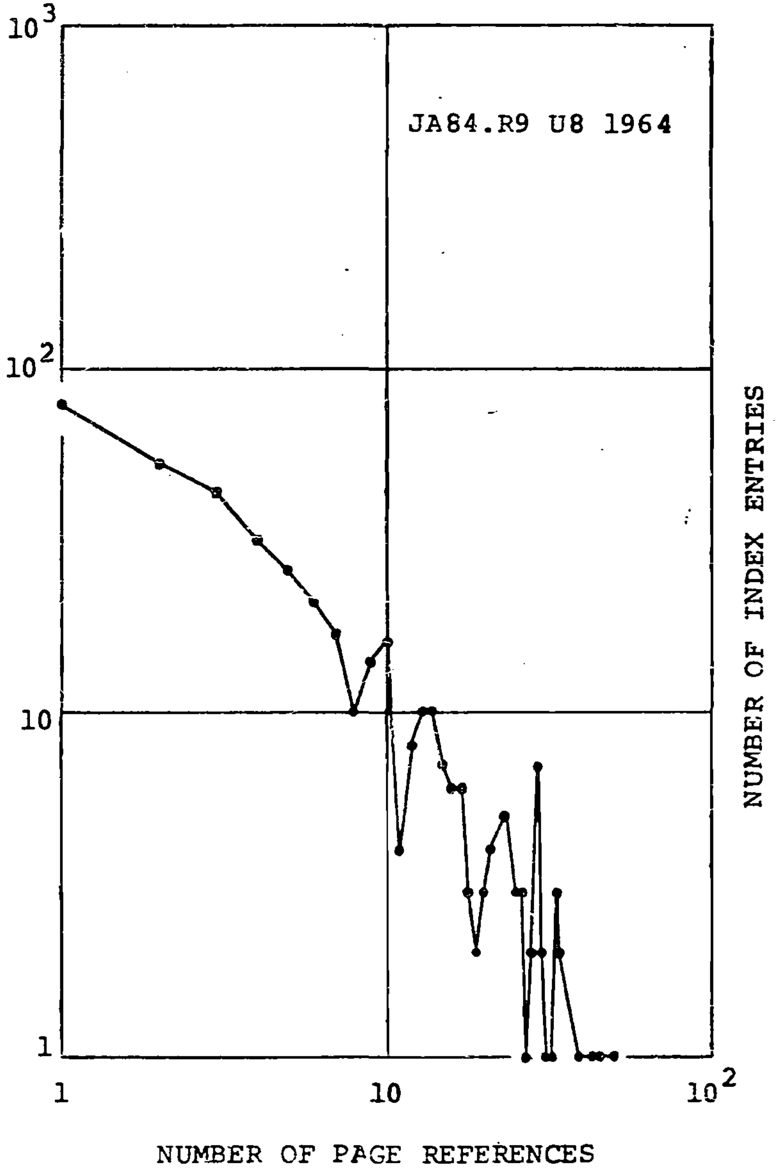

E178.6.S3 1965
10^4
10^3
10^2
10
1
1
10
10^2
NUMBER OF PAGE REFERENCES
NUMBER OF INDEX ENTRIES

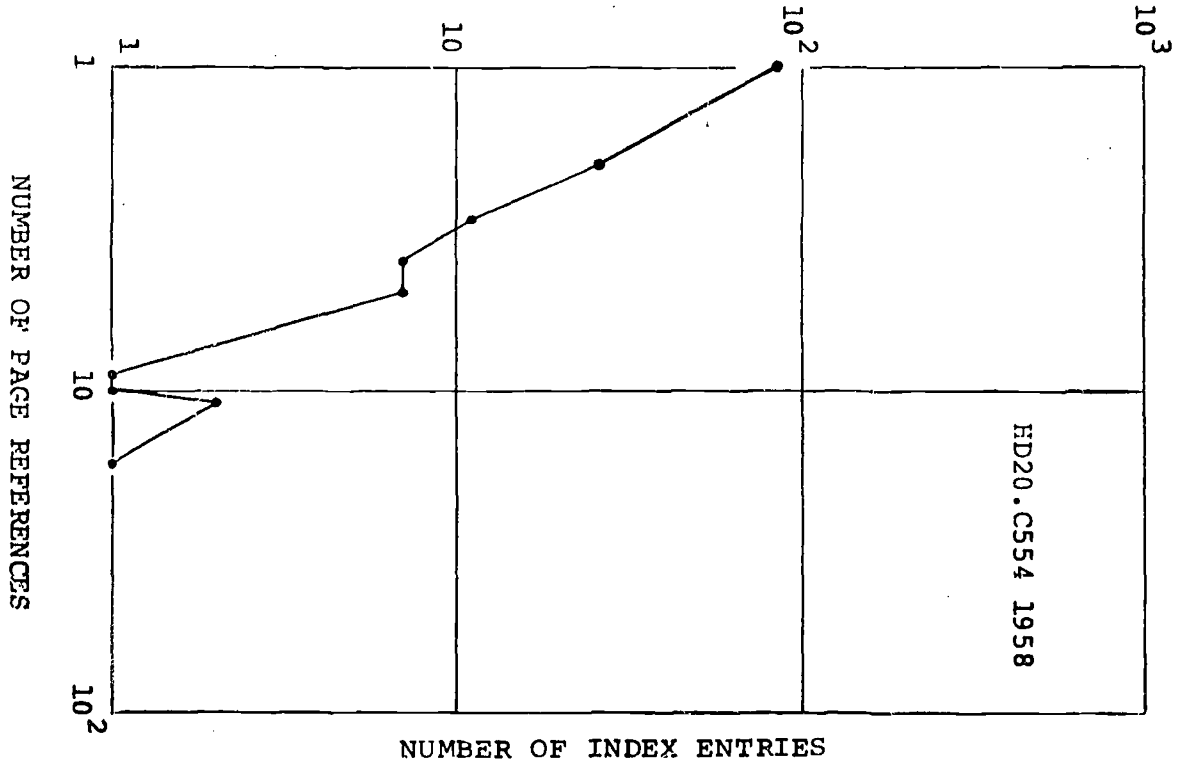


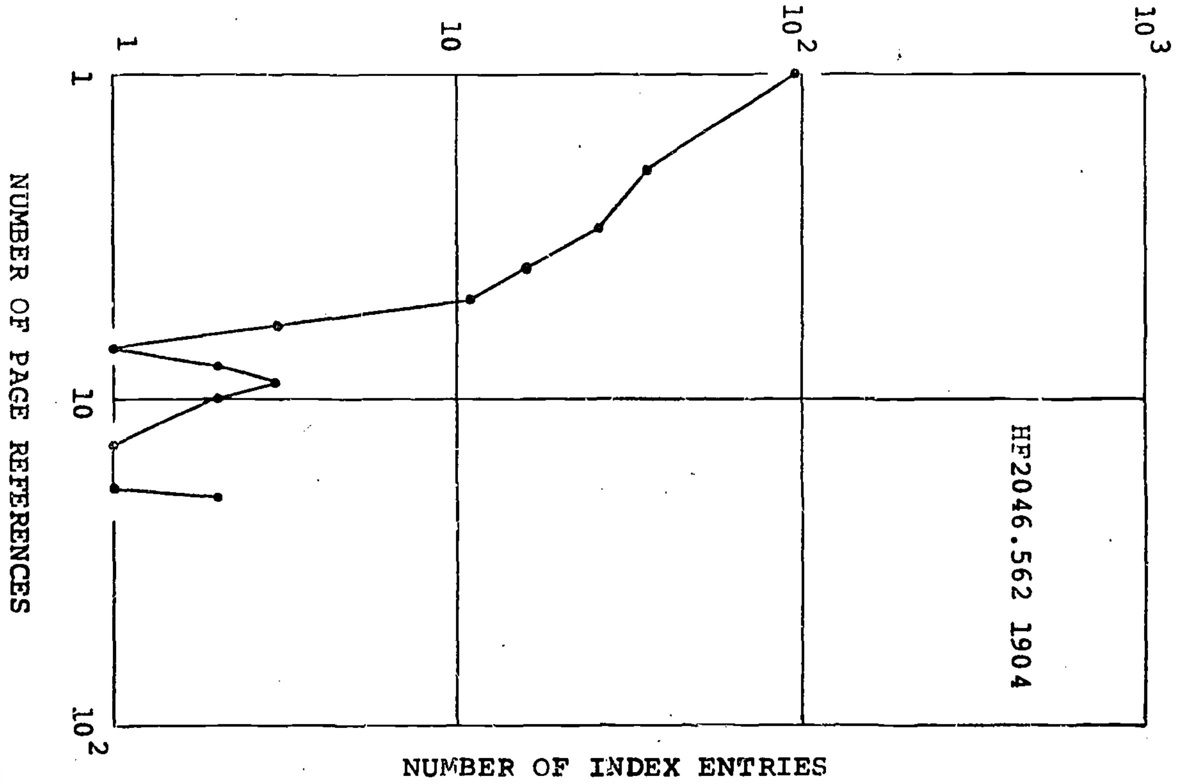
DF521.S4 1966
10^4
10^3
10^2
10
1
1
10
10^2
NUMBER OF INDEX ENTRIES
NUMBER OF PAGE REFERENCES

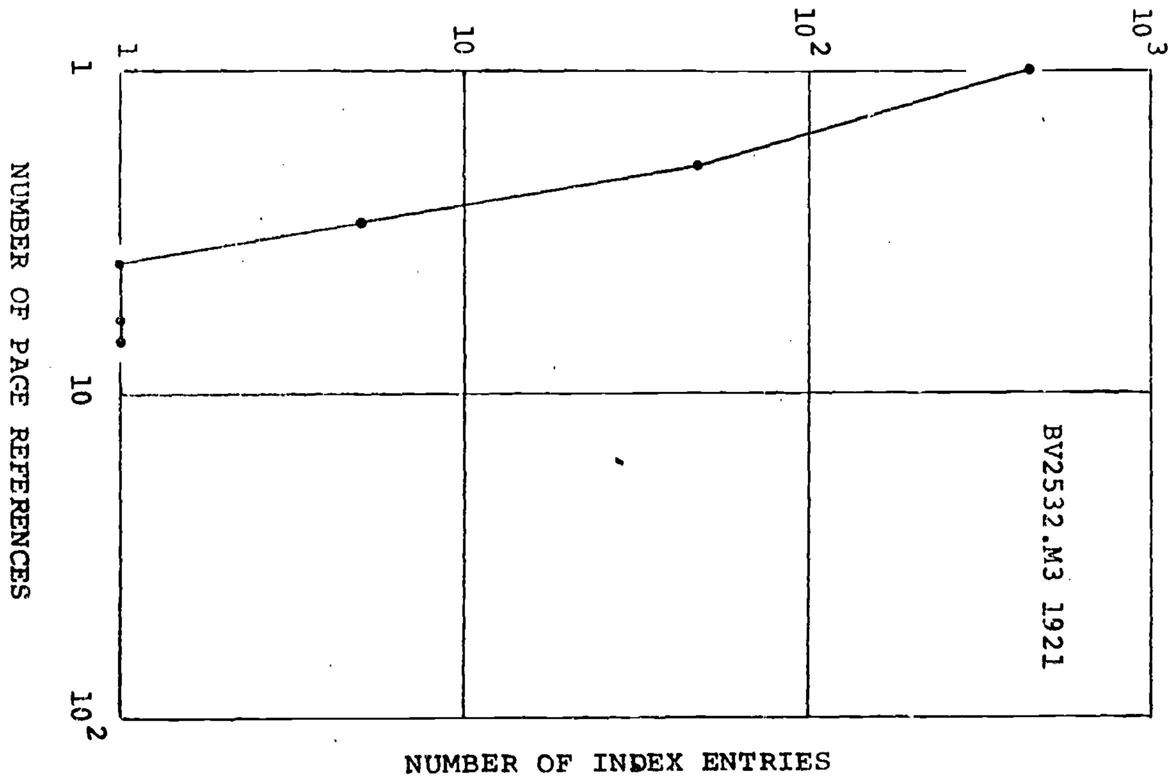

DS423.C85 V4 1953-58
10^4
10^3
10^2
10
1
10^4
10^3
10^2
10
1
10
10^2
NUMBER OF INDEX ENTRIES
NUMBER OF PAGE REFERENCES

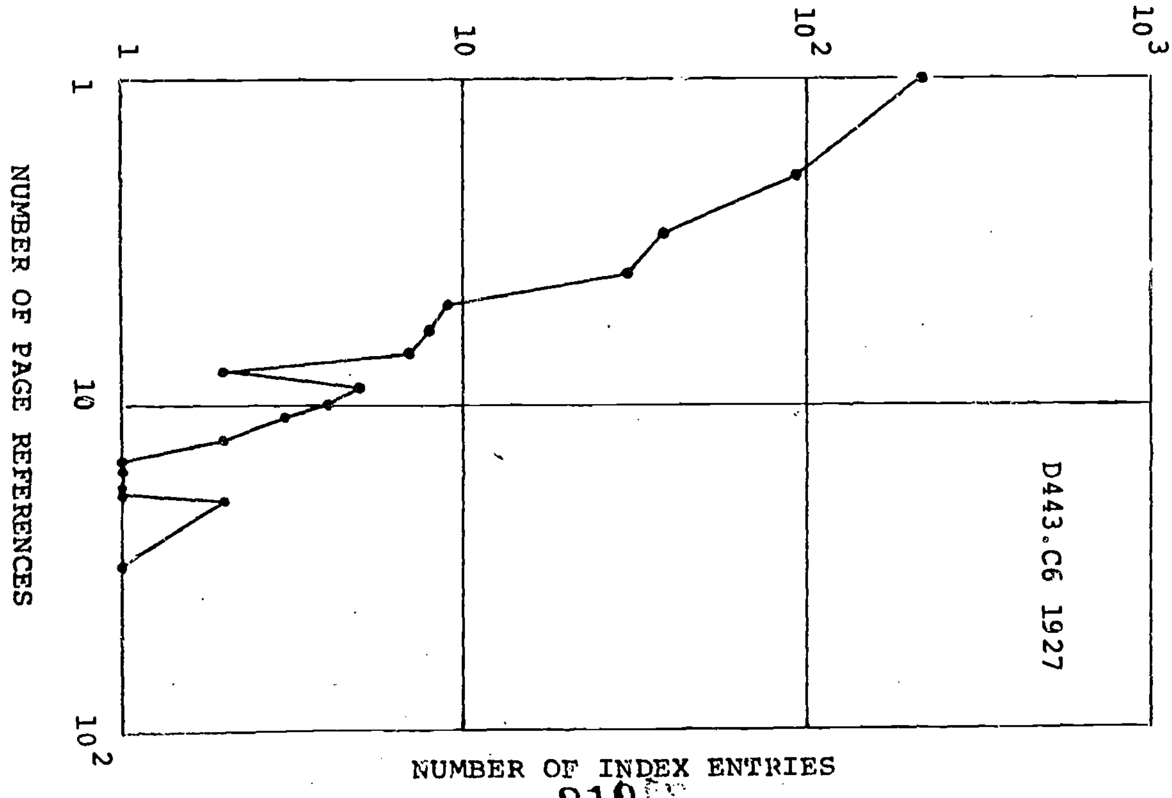

HM66.C7 1920
10^3
10^2
10
1
1
10
10^2
NUMBER OF PAGE REFERENCES
NUMBER OF INDEX ENTRIES

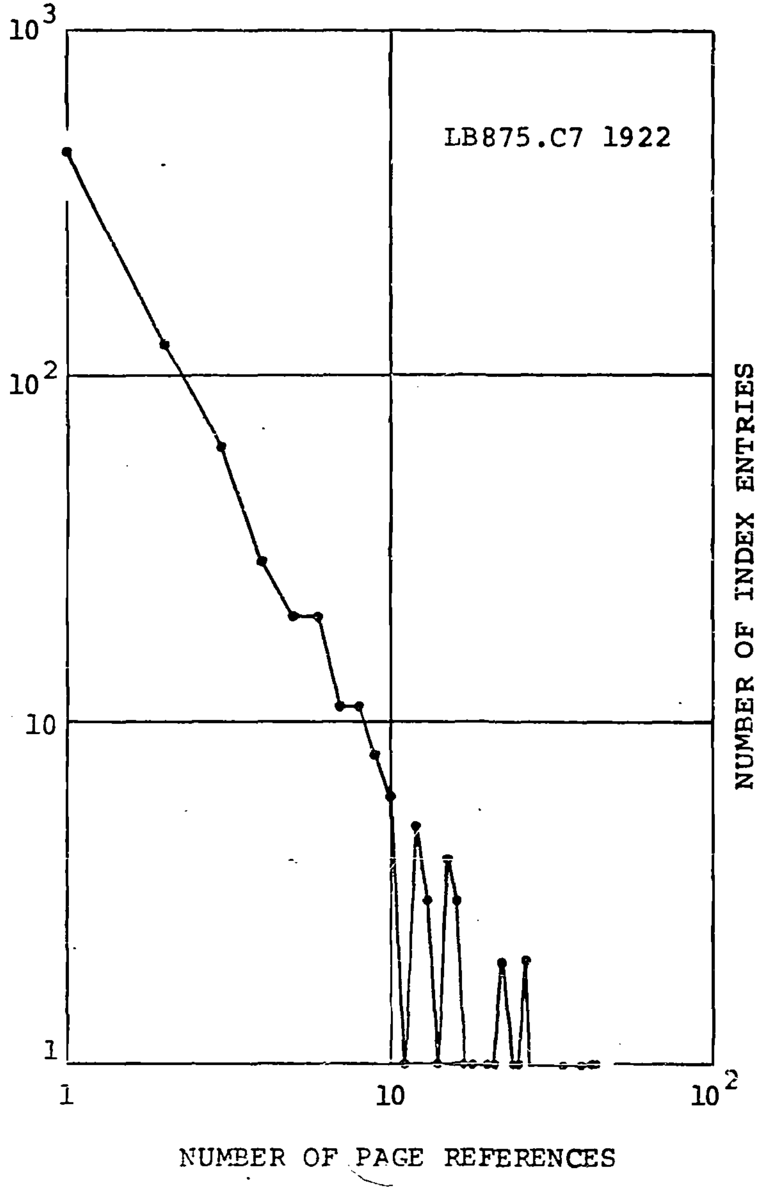

JA84.R9 U8 1964
10^3
10^2
10
1
1
10
10^2
NUMBER OF PAGE REFERENCES
NUMBER OF INDEX ENTRIES

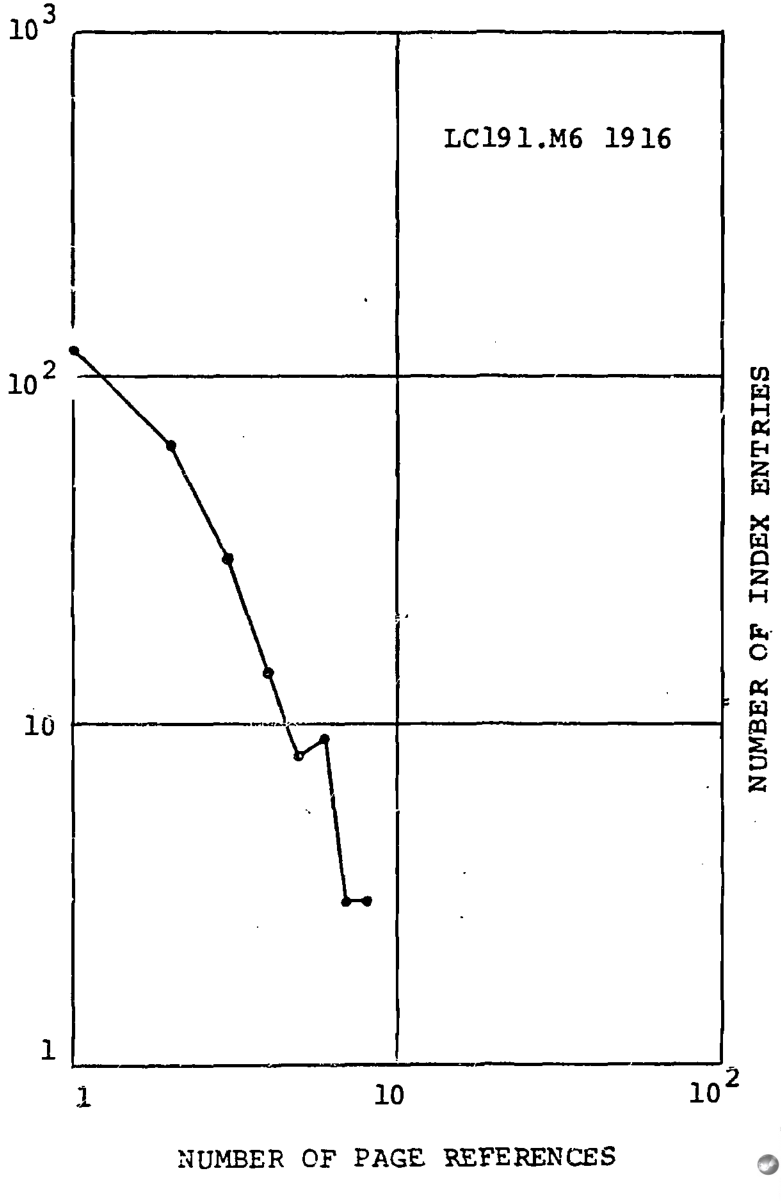

HD20.C554 1958
1
10
10^2
10^3
1
10
10^2
NUMBER OF PAGE REFERENCES
NUMBER OF INDEX ENTRIES

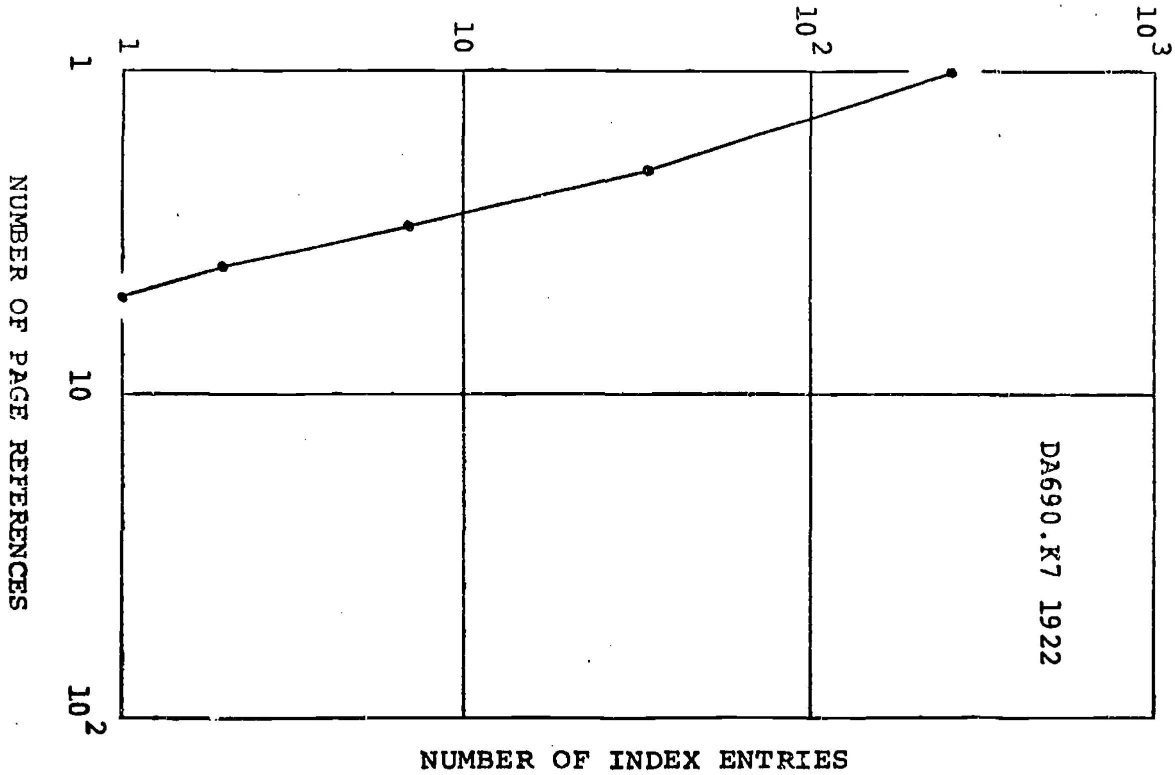

HF2046.562 1904
1
10
10^2
10^3
1
10
10^2
NUMBER OF PAGE REFERENCES
NUMBER OF INDEX ENTRIES

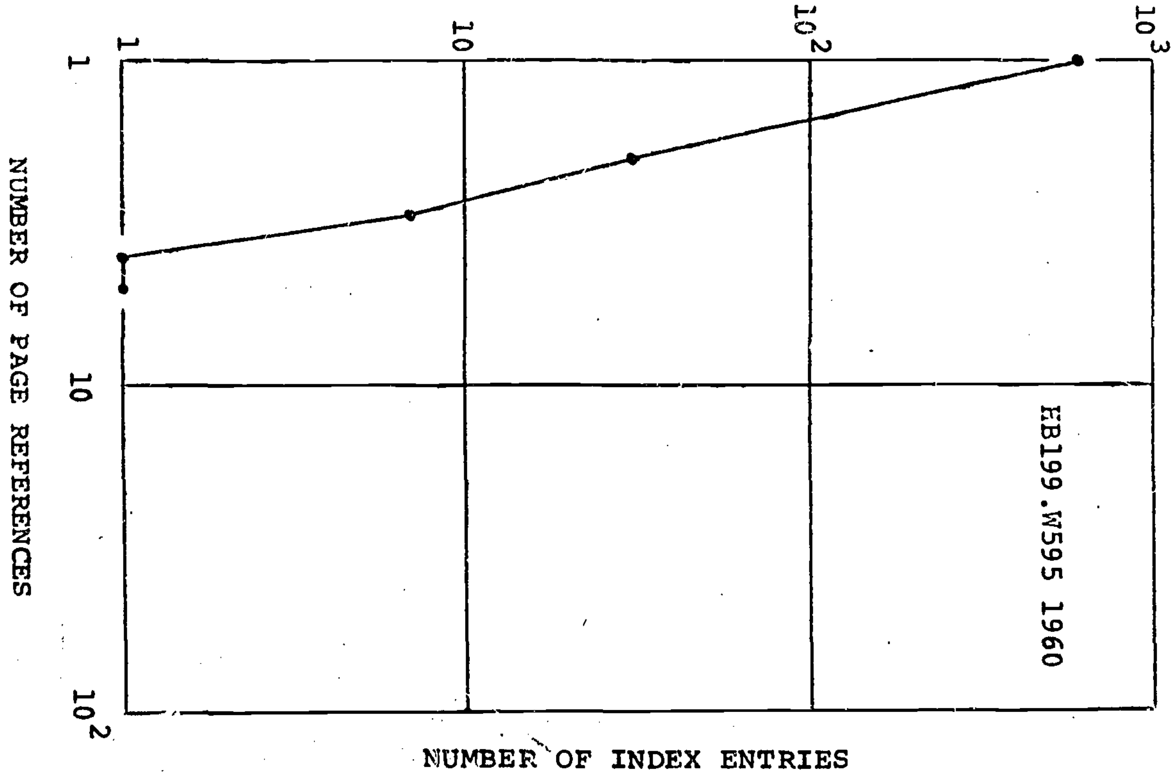

BV2532.M3 1921
NUMBER OF PAGE REFERENCES
NUMBER OF INDEX ENTRIES
1
10
10^2
10^3

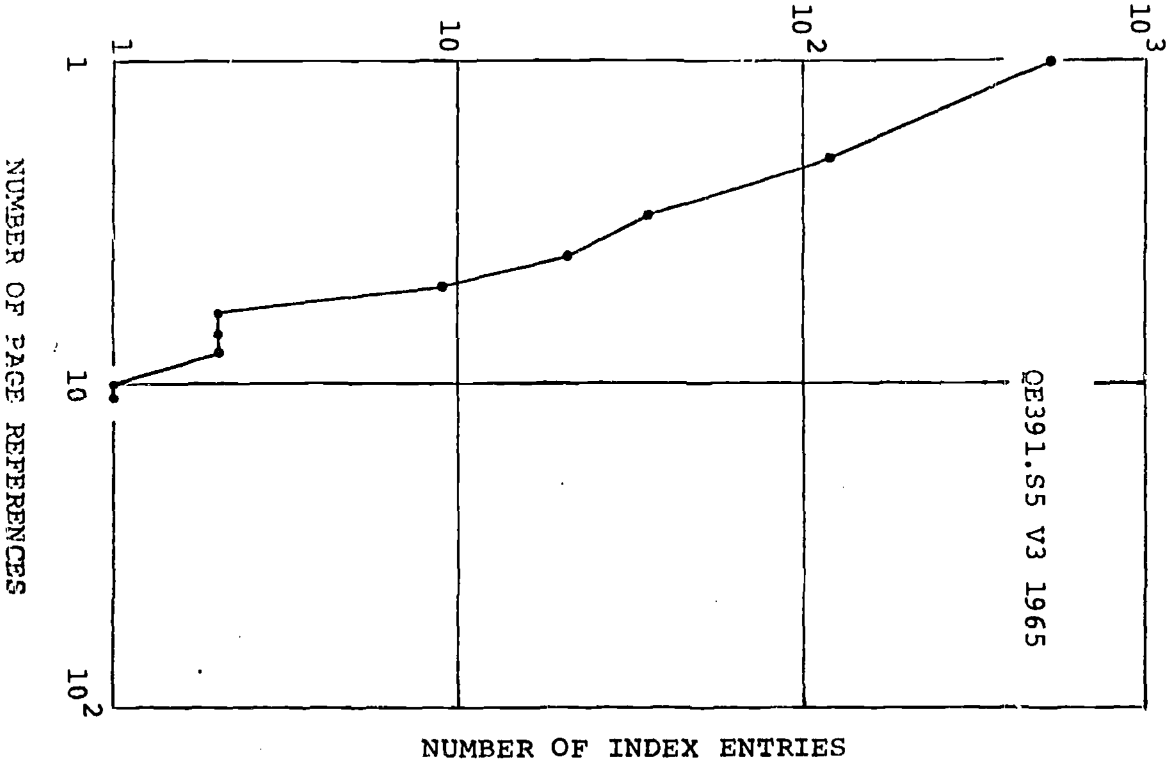

D443.C6 1927
NUMBER OF PAGE REFERENCES
NUMBER OF INDEX ENTRIES
1
10
10^2
10^3

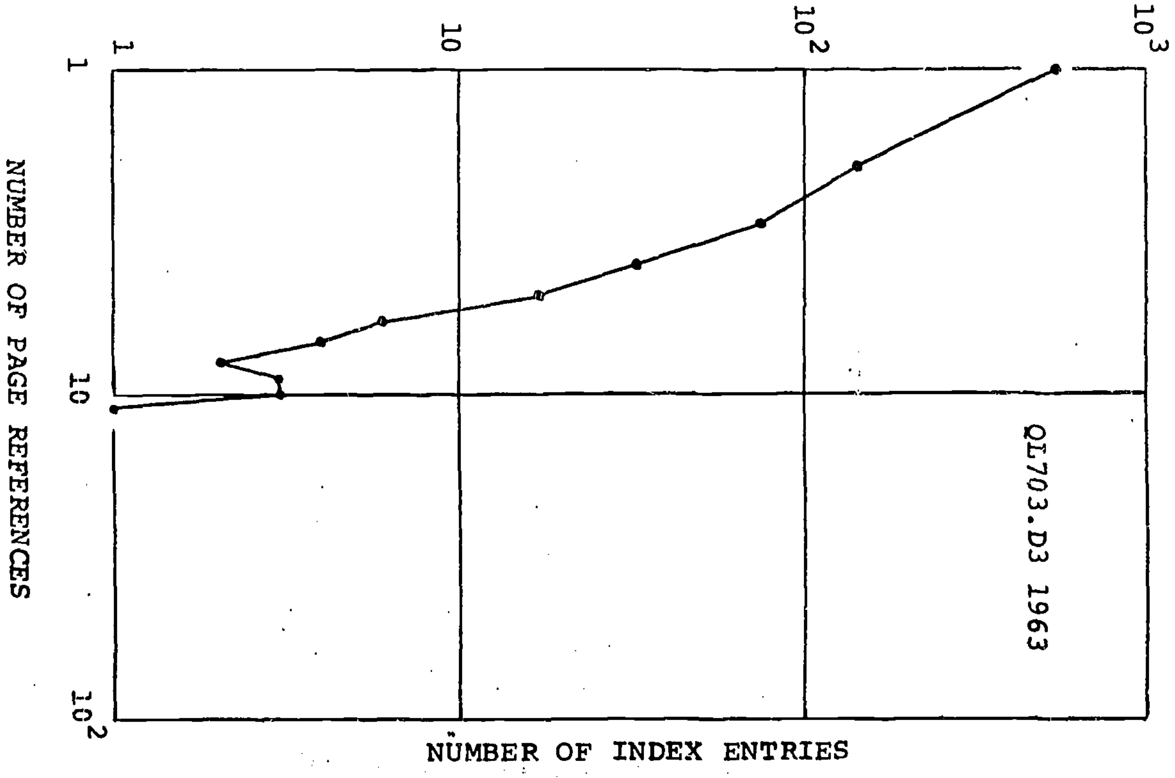

LB875.C7 1922
NUMBER OF INDEX ENTRIES
NUMBER OF PAGE REFERENCES
10³
10²
10
1
1
10
10²

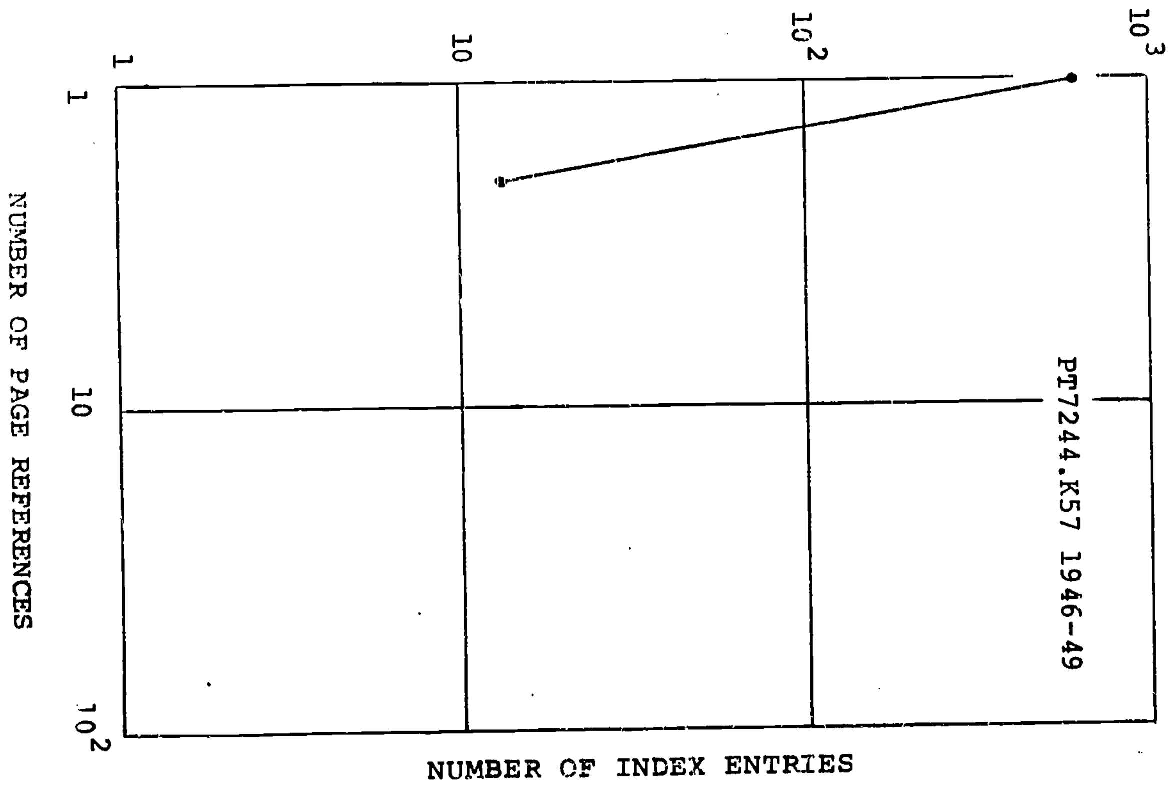

LC191.M6 1916
NUMBER OF INDEX ENTRIES
NUMBER OF PAGE REFERENCES
10³
10²
10
1
1
10
10²

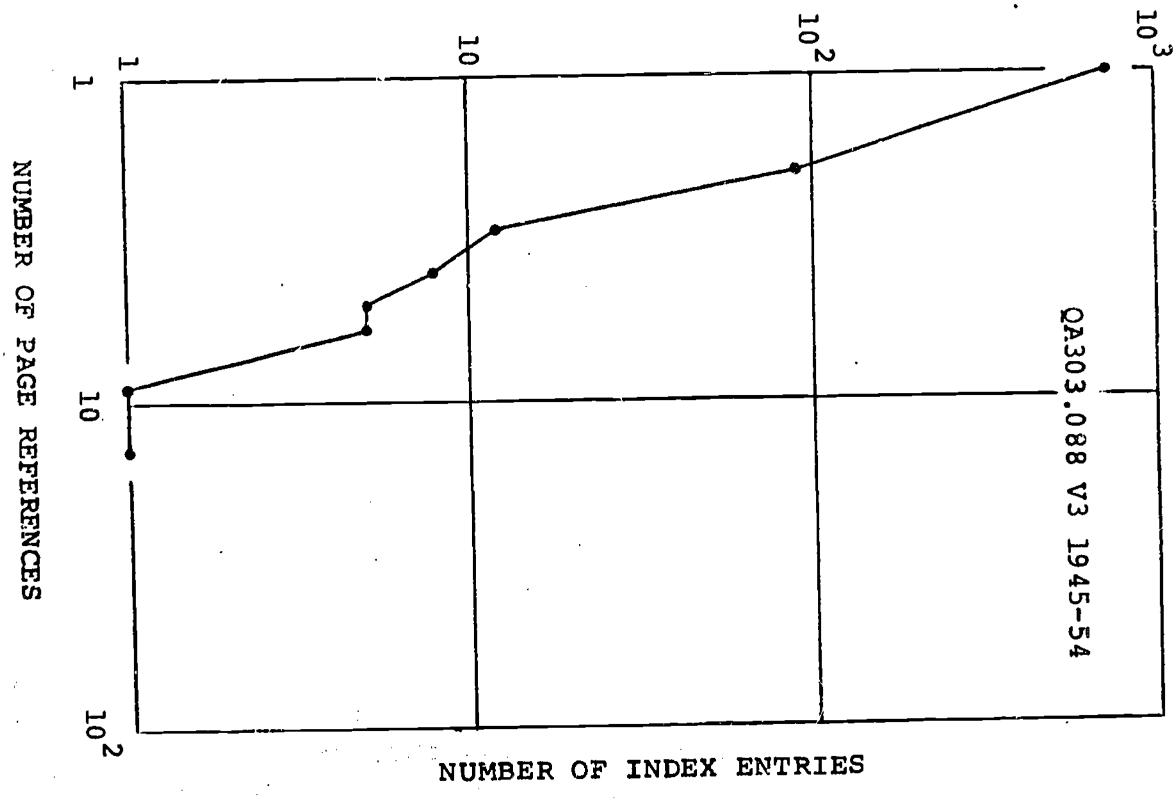

1
10
10^2
10^3
1
10
10^2
NUMBER OF PAGE REFERENCES
DA690.K7 1922
NUMBER OF INDEX ENTRIES

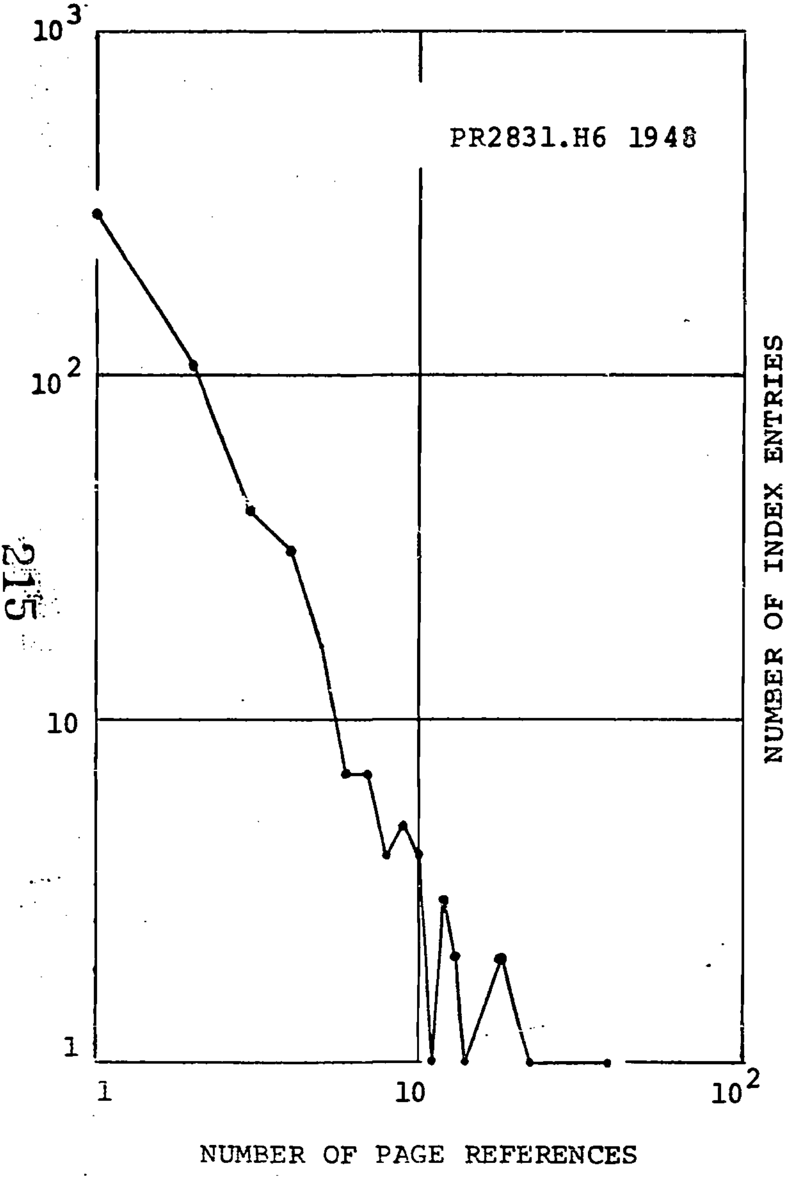

1
10
10^2
10^3
1
10
10^2
NUMBER OF PAGE REFERENCES
EB199.W595 1960
NUMBER OF INDEX ENTRIES

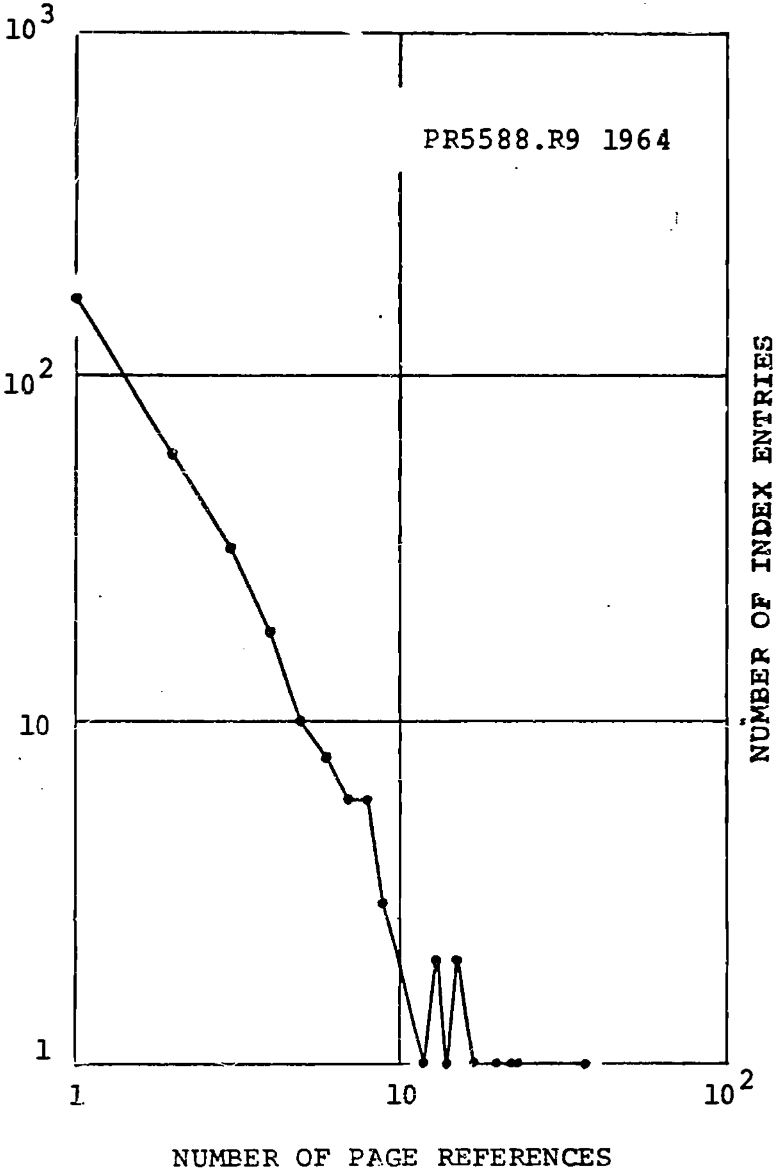

QE391.S5 V3 1965
NUMBER OF INDEX ENTRIES
NUMBER OF PAGE REFERENCES
1
10
10^2
10^3

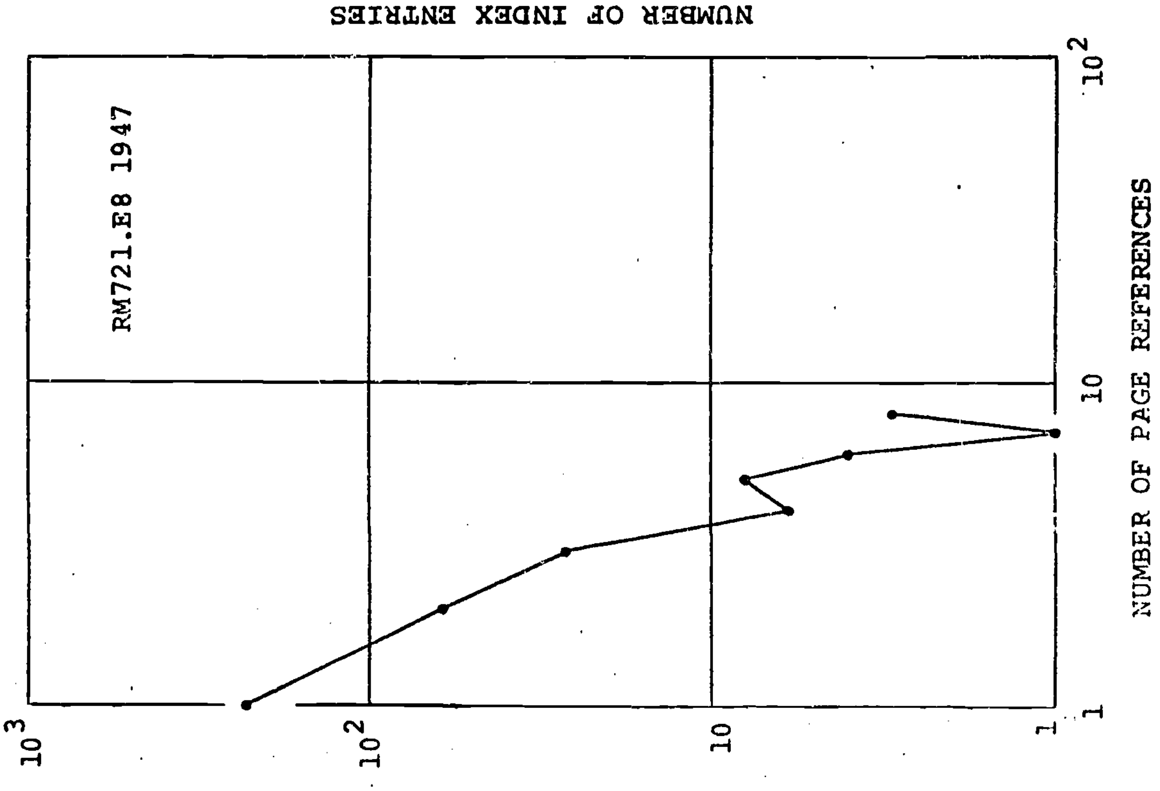

QL703.D3 1963
NUMBER OF INDEX ENTRIES
NUMBER OF PAGE REFERENCES
1
10
10^2
10^3


PT7244.K57 1946-49
NUMBER OF PAGE REFERENCES
1
10
10^2
NUMBER OF INDEX ENTRIES
1
10
10^2
10^3


QA303.088 V3 1945-54
NUMBER OF PAGE REFERENCES
1
10
10^2
NUMBER OF INDEX ENTRIES
1
10
10^2
10^3


PR2831.H6 1948
10^3
10^2
10
1
1
10
10^2
NUMBER OF INDEX ENTRIES
NUMBER OF PAGE REFERENCES


PR5588.R9 1964
10^3
10^2
10
1
1
10
10^2
NUMBER OF INDEX ENTRIES
NUMBER OF PAGE REFERENCES


RM721.E8 1947
NUMBER OF INDEX ENTRIES
NUMBER OF PAGE REFERENCES
10^3
10^2
10
1
1
10
10^2

APPENDIX III

AMALGAMATED ALGORITHMIC INDEX TO ABSTRACTS IN STATISTICS

INDEX TO ABSTRACTS

FROM SELECTED VOLUMES OF THE ANNALS OF MATHEMATICAL STATISTICS

A

- ABSOLUTE CENTRAL MOMENT, 40-0721
- ABSOLUTE CENTRAL MOMENT OF ORDER, 40-0721
- ABSOLUTE CONTINUOUS DISTRIBUTION, 40-0723
- ABSOLUTE DIFFERENCE, 40-0724
- ABSOLUTE FREQUENCIES OF ATTRIBUTES, 33-1480
- ADDITIVITY OF EFFECTS, 34-0357
- ADDITIVITY OF MAIN EFFECTS, 34-0357
- ADMISSIBILITY
  - PROOF OF, 40-1859
- ALMOST INVARIANT CONFIDENCE PROCEDURES, 33-1480
- ALMOST NEGLIGIBLE, 34-0355
- ALTERNATIVE DEFINITION, 41-0329
- ALTERNATIVE PROOF, 40-2219
- ALTERNATIVES
  - PARAMETRIC CLASSES OF, 41-0329
  - SEQUENCE OF, 40-0722
- ANALOGOUS DISTANCE, 40-2219
- ANALYSIS
  - VALIDITY OF, 34-0357
- ANALYSIS OF FATIGUE LIFE, 33-1502
- ANALYSIS OF INFORMATION, 40-0724
- ANALYSIS OF VARIANCE IDENTITY, 33-1480
- ANALYSIS OF VARIANCE TESTS, 34-0357
- ANSCOMBE'S THEOREM
  - EXTENSION OF, 40-2217
- APPLICATIONS OF WEAK CONVERGENCE, 40-2217
- APPROPRIATE DEGREES OF FREEDOM, 40-0721
- APPROXIMATE INTERVALS, 40-1860
- APPROXIMATE POWERS, 40-0722
- ARRIVAL BATCH, 41-0328
- ARY SCALE, 34-0355
- ASSOCIATION
  - MEASURES OF, 33-1480
- ASSUMPTION OF EQUAL, 40-1857
- ASYMPTOTIC BEHAVIOR, 33-1502, 40-1856
- ASYMPTOTIC COMPARISONS, 40-1858
- ASYMPTOTIC COVARIANCE MATRIX, 40-1858
- ASYMPTOTIC DISTRIBUTION, 40-0723, 40-1858
- ASYMPTOTIC DISTRIBUTION THEORY, 40-0721
- ASYMPTOTIC EXPANSIONS, 40-0722
- ASYMPTOTIC NORMALITY, 41-0329
- ASYMPTOTIC NORMALITY OF MAXIMUM LIKELIHOOD
- ASYMPTOTIC OPTIMALITY, 40-1857
- ASYMPTOTIC OPTIMUM PROPERTIES, 40-0722
- ASYMPTOTIC PROPERTIES, 34-0355, 40-1857
- ASYMPTOTIC TESTS, 40-0720
- ASYMPTOTIC THEOREMS
  - DERIVATION OF, 40-2217
- ASYMPTOTIC THEOREMS IN STATISTICS, 40-2217
- ASYMPTOTIC THEORY OF LINEAR COMBINATIONS, 40-2217
- ATTRIBUTE SEQUENCES, 33-1480
- ATTRIBUTES
  - ABSOLUTE FREQUENCIES OF, 33-1480
- AUTHOR USES, 34-0358
- AUXILIARY VARIATE, 40-2219
- AVERAGE OF DISTINCT UNITS, 34-0357
- AVERAGE RISK CRITERION, 40-2219
- AVERAGE SAMPLE SIZE, 40-2216

B

- BAHADUR EFFICIENCY, 40-1858
- BARTLETT'S TEST, 40-0722
- BASIS OF SAMPLES, 40-2216
- BATCH ARRIVAL DISTRIBUTION, 41-0328
- BATCH SERVICE, 41-0328
- BAYES RULE, 40-0720
- BESSEL FUNCTIONS OF MATRIX ARGUMENT, 34-0358
- BEST INVARIANT, 40-1856
- BIVARIATE CASE, 40-0724
- BIVARIATE GAUSSIAN DENSITY FUNCTION, 40-0724
- BLYTH
  - METHODS OF, 40-1859
- BODY OF NORMAL-THEORY TECHNIQUES, 40-1858
- BOREL S-ALGEBRA, 41-0330
- BOUNDS
  - EXISTENCE OF, 41-0329
- BOX'S IDEA, 40-1858

C

- CANONICAL CORRELATIONS, 40-0724
- CASE OF PHASE SERVICE TIME DISTRIBUTION, 41-0328
- CAUCHY DISTRIBUTION, 40-0723
- CELL ENTRIES, 40-0724
- CELL MEANS, 34-0357
- CHARACTERISTIC FEATURE, 33-1480
- CHARACTERISTIC FUNCTION, 40-1860
- CHARACTERISTIC LIFE, 33-1502
- CHARACTERIZATIONS OF SYMMETRIC STABLE PROCESSES, 40-1859
- CHARACTERIZE DEPARTURES, 34-0357
- CHERNOFF-SAVAGE STATISTICS, 40-2217
- CHI-SQUARE DISTRIBUTION, 40-0721
- CHI-SQUARE VARIATE, 40-1860
- CLASS OF CONJUGATE PRIOR DISTRIBUTIONS, 41-0328
- CLASS OF CONJUGATE PRIORS, 41-0328
- CLASS OF NON-PARAMETRIC ALTERNATIVES, 41-0329
- CLASS OF NORMAL DISTRIBUTIONS, 33-1506
- CLASS OF PROCESSES, 40-1856
- CLASS OF SEQUENTIAL PROCEDURES, 40-2216
- CLASS OF STOCHASTIC PROCESSES, 40-1856
- CLASSES OF DISTRIBUTIONS, 33-1506
- CLASSES OF ESTIMATORS, 41-0329
- CLASSIFICATION PROCEDURES, 34-0358
- COEFFICIENTS OF VARIATION, 40-0723
- COLUMN VECTORS, 34-0356
- COMBINATORIAL METHODS, 40-1856
- COMPACT SETS, 41-0330
- COMPARATIVE STUDY, 34-0355
- COMPARE BARTLETT'S TEST, 40-0722
- COMPLETE INVARIANT, 40-2219
- COMPLETE SAMPLES, 40-0723
- COMPLEX CASE, 40-0721
- COMPONENTS IN MODEL II ANOVA, 40-0720
- COMPOUND POISSON PROCESS, 40-1856
- COMPUTATIONAL EXPENDITURES, 40-2216
- CONCEPTS OF PITMAN EFFICIENCY, 40-1859
- CONDITIONAL DISTRIBUTION, 40-2217
- CONDITIONS OF WALD, 40-1857
- CONFIDENCE INTERVAL, 40-0720, 40-1860
- CONFIDENCE PROCEDURES MEASURABLE, 33-1480
- CONFIDENCE REGION, 33-1480, 40-2216, 40-2217
- CONFIDENCE SETS, 40-1860
- CONJUGATE PRIOR DISTRIBUTIONS
  - CLASS OF, 41-0328
- CONJUGATE PRIORS
  - CLASS OF, 41-0328
- CONSERVATIVE NONPARAMETRIC TEST, 40-1857
- CONSISTENCY PROPERTIES, 40-1856
- CONSTANT MULTIPLE, 40-1859
- CONSTANT SCATTER, 40-1859
- CONSTANT TIMES, 40-1860
- CONTINGENCY TABLE, 33-1480, 40-0724
- CONTINOUS MEASURABLE FUNCTION, 41-0329
- CONTINUOUS CDF'S, 34-0357
- CONTINUOUS FUNCTION, 40-0722, 40-1859
- CONTINUOUS POPULATIONS, 40-2216
- CONTINUOUS PROBABILITY DENSITY FUNCTION, 34-0355
- CONTINUOUS TIME VERSION OF FELLER'S COMBINATORIAL LEMMA, 40-1856
- CONVENTIONAL HISTOGRAM, 40-1856
- CONVERGENCE
  - RAPIDITY OF, 40-0723
  - RATE OF, 40-0722
- CONVERGENCE IN DISTRIBUTION, 41-0330
- CONVERGENCE IN PROBABILITY, 40-1859
- CONVEX LOSS FUNCTION, 40-1859
- CONVOLUTION TECHNIQUES, 40-0721
- CORRELATION COEFFICIENT, 40-0720, 40-0724, 40-1859
- CORRELATION PROBLEM, 41-0329
- COVARIANCE
  - UNIVARIATE ANALYSIS OF, 40-0721
- COVARIANCE MATRICES, 40-1858
  - EQUALITY OF, 40-1858
  - STRUCTURE OF, 40-1850
- COVARIANCE MATRIX, 34-0358, 40-0722, 40-1856, 40-1857, 40-2216
- COVERAGE PROBABILITY, 40-1857

D

- DAM MODEL, 40-1856
- DECIMAL PLACES, 40-0723
- DECISION FUNCTIONS, 40-0722
- DECISION RULES, 34-0355
- DEFICIENCIES W.R.T, 40-2219
- DEFINITION OF LINEAR SUFFICIENCY, 41-0329
- DEGREES OF FREEDOM, 40-0720, 40-0722
- DEPENDENT GAUSSIAN VECTORS, 40-0724
- DEPENDENT VARIABLE REPRESENTS DEVIATIONS, 34-0357
- DERIVATION OF ASYMPTOTIC THEOREMS, 40-2217
- DES RAJ, 34-0357
- DESIGN MATRIX, 40-0722
- DIAGRAM OF SERIAL ASSOCIATION, 33-1480
- DIFFERENCE ESTIMATOR, 40-2219
- DIMENSIONAL DISTRIBUTION, 34-0355
- DIMENSIONAL VECTOR, 34-0358
- DIMENSIONS
  - FLAT SPACE OF, 40-0720
- DISCRETE RANDOM VARIABLES
  - SET OF, 40-2217

DISJOINT INTERVALS, 40-2219
DISTANCES
  ESTIMATORS OF, 40-0722
DISTINCT UNITS
  AVERAGE OF, 34-0357
DISTRIBUTION
  CONVERGENCE IN, 41-0330
  RANDOM VARIABLE INVARIANT IN, 40-2216
DISTRIBUTION FUNCTION, 40-0721, 40-1857, 40-2216, 40-2217, 41-0329
DISTRIBUTION-FREE INFERENCES, 34-0355
DISTRIBUTION-FREE TOLERANCE-LIMIT TABLES
  EXTENSIVE SET OF, 34-0355
DISTRIBUTIONS
  CLASSES OF, 33-1506
DISTRIBUTIONS OF POINTS, 33-1506
DOUBLE EXPONENTIAL DISTRIBUTION, 40-0723
DRAWS
  NUMBER OF, 40-2219

E

EASTERN CONNECTICUT STATE COLLEGE, 40-2217
EFFECTS
  ADDITIVITY OF, 34-0357
EMPIRICAL PROCESS, 40-2217
EMPIRICAL PROCESSES, 41-0330
EQUAL
  ASSUMPTION OF, 40-1857
EQUAL DISTRIBUTIONS
  HYPOTHESIS OF, 34-0355
EQUAL SIZE
  SUCCESSIVE CLUSTERS OF, 33-1480
EQUALITY
  HYPOTHESIS OF, 40-1858
EQUALITY OF COVARIANCE MATRICES, 40-1858
EQUALITY OF VARIANCES, 40-1858
ERROR LOSS FUNCTION, 40-1859
ERROR MODEL, 40-2216
ERRORS OF MISCLASSIFICATION, 40-2216
ERRORS OF PREDICTIONS, 34-0358
ESTIMABLE FUNCTION, 40-0721
ESTIMATORS
  CLASSES OF, 41-0329
ESTIMATORS OF DISTANCES, 40-0722
EUCLIDEAN N-SPACE, 40-2218, 40-2219
EXACT CONFIDENCE INTERVALS, 40-1860
EXISTENCE OF BOUNDS, 41-0329
EXPECTATION IN THEOREM, 41-0328
EXPONENTIAL DISTRIBUTION, 40-2219
EXPONENTIAL FAMILIES, 40-1860
EXPONENTIAL RANDOM VARIABLES
  RANDOM SUM OF, 41-0328
EXTENSION OF ANSCOMBE'S THEOREM, 40-2217
EXTENSIVE SET OF DISTRIBUTION-FREE TOLERANCE-LIMIT TABLES, 34-0355
EXTREME CASE, 33-1480

F

FACTORS
  MAXIMUM NUMBER OF, 40-0720, 40-0723, 40-2217, 40-2218
FAILURE TIMES, 33-1502
FAMILIES OF FINITE MEASURES, 40-2219
FAMILY OF PROBABILITY DENSITIES, 40-0722
FAMILY OF PROBABILITY [illegible], 40-0722
FATIGUE LIFE
  ANALYSIS OF, 33-1502
FELLER'S COMBINATORIAL LEMMA
  CONTINUOUS TIME VERSION OF, 40-1856
FINDINGS OF HAJEK, 40-1857
FINITE MEASURE, 40-0722
FINITE MEASURE FIELD, 33-1480
FINITE MEASURES
  FAMILIES OF, 40-2219
FINITE MOMENTS, 40-0721
FINITE NUMBER, 40-2216, 40-2217
FINITE POPULATION, 40-2217, 40-2218, 40-2219
FINITE PROJECTIVE GEOMETRY, 40-0723
  POINTS IN, 40-0720
FINITE PROJECTIVE SPACE
  POINTS IN, 40-2217, 40-2218
FINITE SEQUENCES
  RANDOMNESS IN, 33-1480
FISHER-YATES TEST, 41-0329
FIXED-SAMPLE PROCEDURE, 40-2216
FLAT SPACE OF DIMENSIONS, 40-0720
FORM
  FUNCTIONS OF, 40-1860
  HYPOTHESES OF, 40-1860
FORM OF GUESS, 40-0722
FORM REPRESENTATIONS, 40-0721
FORMAL PROCESS, 40-1860
FORMAL SOLUTIONS, 34-0358
FORMULAE OF INTRACLASS CORRELATION, 33-1480
FREEDOM
  APPROPRIATE DEGREES OF, 40-0721
  DEGREES OF, 40-0720, 40-0722
  HYPOTHESIS DEGREES OF, 40-0721
  N-2 DEGREES OF, 40-2220
FULL SET, 40-1858
FUNCTIONS OF FORM, 40-1860
FUTURE OBSERVATION, 34-0356

G

GALOIS FIELD, 40-0720, 40-2217, 40-2218
GALOIS FIELD GF(5), 40-0723
GALTON TEST, 40-2217
GAMMA TEST, 34-0355
GAUSS-MARKOV ESTIMATE, 40-1860
GAUSS-MARKOV THEOREM, 40-0723
GAUSSIAN PROCESSES, 41-0330
GENERALITY
  LOSS OF, 40-1859
GODAMBE'S DEFINITION OF LINEAR SUFFICIENCY, 41-0329
GRAND MEAN PLUS, 34-0357
GUESS
  FORM OF, 40-0722

H

HAGA TEST, 40-2216
HAJEK
  FINDINGS OF, 40-1857
HAND COMPUTATION, 40-1857
HAROLD CRAMER VOLUME, 40-1856
HARTER-LUM TECHNIQUES, 34-0357
HIGHEST ORDER INTERACTION, 34-0357
HISTOGRAMS
  TYPES OF, 40-1856
HOGBEN ET AL, 40-0720
HOMOGENEITY
  TESTS OF, 40-0724
HOMOGENEITY OF VARIANCES, 40-0722
HOMOGENEOUS STOCHASTIC PROCESS, 40-1859
HORVITZ-THOMPSON ESTIMATOR, 40-2218, 41-0329
HYPERGEOMETRIC DENSITY, 40-0720
HYPOTHESES
  TESTS OF, 34-0358
HYPOTHESES OF FORM, 40-1860
HYPOTHESIS DEGREES OF FREEDOM, 40-0721
HYPOTHESIS OF EQUAL DISTRIBUTIONS, 34-0355
HYPOTHESIS OF EQUALITY, 40-1858
HYPOTHESIS OF MARGINAL HOMOGENEITY, 40-0724

I

IDEA OF KATTI, 40-0722
IDENTICAL SYMMETRIC DENSITIES
  TRANSLATION PARAMETERS OF, 40-1859
IDLE SERVERS
  NUMBER OF, 41-0328
INADMISSIBLE ESTIMATOR, 40-1859
INCLUSION PROBABILITIES, 40-2218
  SEQUENCE OF, 40-2218
INCLUSION PROBABILITY, 40-2218
INCREASE
  POINT OF, 34-0355
INDEPENDENCE OF SETS, 40-1858
INDEPENDENT COLLECTION, 34-0355
INDEPENDENT ESTIMATE, 40-1860
INDEPENDENT IN PAIRS, 34-0355
INDEPENDENT INCREMENTS, 40-1859
INDEPENDENT NORMAL RANDOM VARIABLES, 40-1859
INDEPENDENT OBSERVATIONS
  SEQUENCE OF, 40-2217
INDEPENDENT P-VARIATE NORMAL POPULATIONS, 40-1856
INDEPENDENT POISSON PROCESSES, 33-1502
INDEPENDENT PROCESSES, 40-1856
INDEPENDENT RANDOM SAMPLES, 40-2216
INDEPENDENT RANDOM VARIABLES, 34-0357, 40-1857
  SUMS OF, 41-0329
INDEPENDENT SAMPLE, 34-0358
INDEPENDENT UNOBSERVABLE RANDOM VARIABLES, 40-1857
INDEPENDENT VARIABLE, 34-0357, 40-0720
INDICES
  TRIPLET OF, 34-0355
INDIVIDUAL CUSTOMER, 40-1856
INDIVIDUAL REGRESSION COEFFICIENT, 40-1858
INFINITE CASES, 40-0720
INFORMATION
  ANALYSIS OF, 40-0724
  MEASURES OF, 40-2219
INFORMATION- THEORETIC INEQUALITY, 40-0724
INITIAL QUEUES OF SIZES, 40-1856
INTEGERS
  PAIR OF, 34-0355
INTEGRANDS SATISFY, 40-1859
INTERARRIVAL DISTRIBUTION, 41-0328
INTERARRIVAL TIMES, 41-0328
INTEREST
  PROPERTIES OF, 40-0722
  REGION OF, 40-1860
INTRACLASS ASSOCIATION, 33-1480
INTRACLASS CORRELATION
  FORMULAE OF, 33-1480
INVARIANT MARKOV KERNELS, 40-2219
IOWA CITY, 41-0328

J

JACKKNIFE ESTIMATE, 4(
JACKKNIFE PROCEDURE, 4
JACKKNIFE TESTS, 40-18
JOINT DISTRIBUTION, 34
40-2216
JOINT STATISTICAL CONFERENCE, 41-0328

K

K-DECISION PROBLEMS, 4
K-DIMENSIONAL UNIT-CUB
0330
K-STATISTICS
  MULTIPLICATION OF, 4(
KATTI
  IDEA OF, 40-0722
KOLMOGOROV TEST STATIS
41-0330
KOLMOGOROV-SMIRNOV TES
2216

L

LACK OF RANDOMNESS, 33-
LADDER PROCESS, 40-1856
LADDER PROCESSES, 40-18
LARGE NUMBERS
  LAW OF, 40-2218
LARGER MEAN, 40-1859
LATENT ROOTS, 34-0358
LATIN SQUARE, 34-0355
LAW OF LARGE NUMBERS, 4(
2218
LEAST SQUARES, 33-1502
LEHMANN'S
  VERSION OF, 34-0357
LEHMANN'S TEST, 40-0722
  NULL DISTRIBUTION OF,
  0722
LEXICOGRAPHIC ORDER, 34-
LIKELIHOOD FUNCTION, 40-
LIKELIHOOD RATIO, 40-072
LIKELIHOOD RATIO CRITERI
40-0722
LIKELIHOOD RATIO CRITERI
40-0721
LIKELIHOOD RATIO TEST, 4
0722
LIKELIHOOD-RATIO TEST, 4
1858
LIMIT THEOREM, 40-1856
LINEAR COMBINATIONS
  ASYMPTOTIC THEORY OF, (
  2217
LINEAR COMBINATIONS OF O(
STATISTICS, 40-2217
LINEAR ESTIMATORS, 41-03
LINEAR FORMS, 34-0355
LINEAR FUNCTION, 40-1860
LINEAR HYPOTHESIS PROBLEM
40-0721
LINEAR MODEL, 34-0357, 40
0722, 40-1857, 40-2220
LINEAR REGRESSION, 40-185
LINEAR SUFFICIENCY, 41-03
  DEFINITION OF, 41-0329
  GODAMBE'S DEFINITION OF
  41-0329
LINEAR SUFFICIENT ESTIMAT
41-0329
LINEARITY VERSUS CONVEXIT
  RANK TEST OF, 40-1857
LOGISTIC DISTRIBUTION, 40
0723
LOGNORMAL DISTRIBUTION
  MEAN OF, 40-1860
LOSS OF GENERALITY, 40-18
LOWER BOUND, 41-0329
LOWER ORDER INTERACTION, (
0720, 40-0723, 40-2217, 4
2218

M

M-WAY MARGINAL HOMOGENEITY
  QUESTION OF, 40-0724

MAIN EFFECT, 34-0357, 40-2217
MAIN EFFECTS
  ADDITIVITY OF, 34-0357
MAIN PORTION, 40-1857
MANN-WHITNEY-WILCOXON TESTS, 34-0355
MENUCA PROBLEM, 40-0721
MARGINAL DENSITIES, 40-0724
MARGINAL DISTRIBUTION, 34-0355, 41-0328
MARGINAL HOMOGENEITY
  HYPOTHESIS OF, 40-0724
  PROBLEM OF, 40-0724
  TESTS OF, 40-0724
MARKOV KERNEL CRITERION, 40-2219
MATRIX ARGUMENT
  BESSEL FUNCTIONS OF, 34-0358
MAXIMAL DISTANCE, 40-2219
MAXIMUM DIAMETER, 40-2217
MAXIMUM INFORMATION EXPERIMENT, 40-2219
MAXIMUM LIKELIHOOD, 40-1856
  METHOD OF, 33-1502
MAXIMUM LIKELIHOOD ESTIMATE, 34-0358
MAXIMUM LIKELIHOOD ESTIMATES
  ASYMPTOTIC NORMALITY OF, 40-2217
MAXIMUM LIKELIHOOD ESTIMATION, 33-1502
MAXIMUM LIKELIHOOD HISTOGRAMS, 40-1856
MAXIMUM NUMBER OF FACTORS, 40-0720, 40-0723, 40-2217, 40-2218
MAXIMUM NUMBER OF POINTS, 40-0720, 40-0723, 40-2217, 40-2218
MCGILL UNIVERSITY, 40-2218, 40-2220
MEAN OF LOGNORMAL DISTRIBUTION, 40-1860
MEAN VECTOR, 40-1857
~~MEAN ZERO, 40-2217~~
MEASURES
  SEQUENCE OF, 41-0330
MEASURES OF ASSOCIATION, 33-1480
MEASURES OF INFORMATION, 40-2219
MEDIAN TEST, 34-0355, 40-2216
MEHLER'S IDENTITY, 40-0724
MELLIN TRANSFORMS, 40-1860
METHOD OF MAXIMUM LIKELIHOOD, 33-1502
METHODS OF BLYTH, 40-1859
MICHIGAN STATE UNIVERSITY, 40-0723
MINIMAX CONFIDENCE PROCEDURES, 33-1480
MINIMAX RISK CRITERION, 40-2219
MINIMUM DISCRIMINATION INFORMATION ESTIMATION
  PRINCIPLE OF, 40-0724
MINIMUM DISCRIMINATION INFORMATION STATISTIC, 40-0724
MINIMUM MEAN SQUARE ESTIMATOR, 41-0329
MINIMUM NUMBER, 40-2217
MINIMUM VARIANCE ESTIMATOR, 40-0721
MIRROR IMAGES, 34-0355
MISCLASSIFICATION
  ERRORS OF, 40-2215
MODEL II ANOVA
  COMPONENTS IN, 40-0720
MONOTONIC FUNCTION, 34-0357
MONTE CARLO, 33-1502
MONTE CARLO POWER COMPARISONS, 40-1857
MULTI-COMPONENT STRUCTURES
  RELIABILITY FUNCTIONS OF, 33-1480
MULTIDIMENSIONAL CONTINGENCY TABLE, 40-0724
MULTIPLE REGRESSION, 34-0357
MULTIPLE REGRESSION OF [illegible]VITY, 34-0357
MULTIPLICATION ALGORITHM, 40-1857
MULTIPLICATION OF K-STATISTICS, 40-1857
MULTIVARIATE CASE, 40-0721, 40-1858
MULTIVARIATE FAMILY, 41-0328
MULTIVARIATE NORMAL LINEAR HYPOTHESIS, 34-0358
MULTIVARIATE REGRESSION, 34-0356
MULTIVARIATE STABLE DISTRIBUTIONS, 40-1860
MULTIVARIATE STATISTICAL ANALYSIS, 40-1860
MULTIVARIATE STATISTICS
  NONCENTRAL DISTRIBUTION PROBLEMS IN, 34-0358

N

N-2 DEGREES OF FREEDOM, 40-2220
NEYMAN ALLOCATION FORMULA, 41-0328
NON RANDOM, 33-1480
NON-CENTRAL BETA VARIABLE, 40-0720
NON-IDENTITY TRANSFORMATIONS
  NUMBER OF, 40-2216
NON-NEGATIVE CONSTANT, 33-1480
NON-NEGATIVE DEFINITE, 40-1860
NON-PARAMETRIC ALTERNATIVES
  CLASS OF, 41-0329
NON-STEADY STATE LAPLACE TRANSFORMS, 40-1856
NON-STOCHASTIC PREDICTORS, 40-1857
NONADDITIVITY
  MULTIPLE REGRESSION OF, 34-0357
NONCENTRAL DISTRIBUTION PROBLEMS IN MULTIVARIATE STATISTICS, 34-0358
NONCENTRAL DISTRIBUTIONS, 34-0358
NONCENTRAL MULTIVARIATE BETA DISTRIBUTION, 34-0358
NONNULL DISTRIBUTIONS, 40-0722
NONPARAMETRIC ALTERNATIVE, 34-0355, 34-0357
NONPARAMETRIC TESTS, 40-2216
NORMAL ALTERNATIVES, 34-0355
NORMAL CASE, 40-1858
NORMAL DISTRIBUTION, 40-0722, 40-0723, 40-1857, 40-1860
NORMAL DISTRIBUTIONS
  CLASS OF, 33-1506
NORMAL MEANS, 40-1859
NORMAL POPULATIONS, 40-0722
NORMAL RANDOM VARIABLE, 40-1860
NORMAL THEORY LIKELIHOOD RATIO STATISTIC, 40-0721
NORMAL THEORY LIKELIHOOD RATIO TEST STATISTIC, 40-0721
NORMAL THEORY T-TEST, 40-1857
NORMAL-THEORY TECHNIQUES
  BODY OF, 40-1858
  ROBUSTNESS OF, 40-1858
NULL DISTRIBUTION OF LEHMANN'S TEST, 40-0722
NULL DISTRIBUTIONS OF WILKS, 40-0721
NULL HYPOTHESIS, 40-0722
NUMBER OF DRAWS, 40-2219
NUMBER OF IDLE SERVERS, 41-0328
NUMBER OF NON-IDENTITY TRANSFORMATIONS, 40-2216
NUMBER OF SERVERS, 41-0328
NUMBER OF SUCCESSES, 40-0T20
NUMBER OF VARIATES, 40-0721
NUMBER SUCCESSES, 40-0720
NUMERICAL COMPARISONS, 34-0357
NUMERICAL EXAMPLES, 40-0722
NUMEROUS EXAMPLES, 40-1860

O

ONE-DIMENSIONAL EMPIRICAL PROCESS CONVERGE, 41-0330
ONTO ITSELF, 33-1480
OPERATIONAL CHARACTERISTICS, 40-2219
OPTIMAL ALLOCATION, 41-0328
OPTIMAL ALLOCATION PROBLEMS, 41-0328
OPTIMAL DESIGN, 40-1858
OPTIMAL HISTOGRAM, 40-1856
OPTIMAL STRATIFIED SAMPLING, 41-0328
OPTIMUM ALLOCATION, 40-2219
OPTIMUM BEST LINEAR, 40-0723
OPTIMUM BLUE'S, 40-0723
OPTIMUM NON-PARAMETRIC STATISTICS, 40-0722
ORDER
  ABSOLUTE CENTRAL MOMENT OF, 40-0721
ORDER ABSOLUTE CENTRAL MOMENT, 40-0721
ORDER STATISTICS, 40-0723
  LINEAR COMBINATIONS OF, 40-2217
  SET OF, 40-0723
OVERALL AVERAGE OF SUBSAMPLE MEANS, 34-0357

P

P-DIMENSIONAL SPACE, 40-221T
P-DIMENSIONAL VARIATE, 41-0328
P-VARIATE DISTRIBUTION, 40-1857
P-VARIATE NORMAL POPULATIONS, 40-2216
PAIR OF INTEGERS, 34-0355
PAIRS
  INDEPENDENT IN, 34-0355
PAPER GENERALIZES, 34-0355
PAPER TREATS, 40-2219
PARAMETER SET, 40-2219
PARAMETRIC CLASSES OF ALTERNATIVES, 41-0329
PAST OBSERVATIONS AVAILABLE, 34-0356
PHASE DISTRIBUTION, 41-0328
PHASE SERVICE TIME DISTRIBUTION
  CASE OF, 41-0328
PITMAN EFFICIENCY, 40-1857
  CONCEPTS OF, 40-1858
PITMAN ESTIMATOR, 40-1859
POINT OF INCREASE, 34-0355
POINT OF VIEW, 41-0328
POINTS
  DISTRIBUTIONS OF, 33-1506
  MAXIMUM NUMBER OF, 40-0720, 40-0723, 40-2217, 40-2218
POINTS IN FINITE PROJECTIVE GEOMETRY, 40-0720
POINTS IN FINITE PROJECTIVE SPACE, 40-2217, 40-2218
POPULATION MEAN, 34-0357, 40-2218
POPULATION SIZE, 40-0720
POPULATION TOTAL, 40-2219, 41-0329
POPULATION VARIANCE, 34-0357
POPULATION VECTOR, 40-2219
POPULATIONS CORRESPOND, 40-1859
POSITION INTERMEDIATE, 34-0355
POSITIVE CONSTANT, 40-1859
POSITIVE DEFINITE, 34-0358
POSITIVE DEFINITE MATRIX, 40-1856
POSITIVE NUMBER, 40-2216
POSITIVE SOLUTION, 40-1859
POSTERIOR COVARIANCE, 41-0328
POWER FUNCTION, 34-0355, 34-0357
POWER FUNCTIONS OF TWO-SAMPLE RANK TESTS, 34-0355
PRE-EMPTIVE RESUME PRIORITY SERVICE DISCIPLINE, 33-1502
PREDICTIONS
  ERRORS OF, 34-0358
PRELIMINARY REPORT, 40-2220
PRELIMINARY TEST, 40-2220
PRINCIPLE OF MINIMUM DISCRIMINATION INFORMATION ESTIMATION, 40-0724
PRIORI KNOWLEDGE, 40-0722
PRIORITY LEVEL, 33-1502
PROBABILISTIC CONVERGENCE, 40-2218
PROBABILISTIC PSEUDO-METRIC SPACE, 40-0722
PROBABILITY
  CONVERGENCE IN, 40-1859
  THEORY OF, 34-0355
PROBABILITY DENSITIES
  FAMILY OF, 40-0722
PROBABILITY DENSITY, 40-1860
PROBABILITY FIELDS, 33-1480
PROBABILITY MEASURES, 40-2219
PROBABILITY OF RANK ORDERS, 34-0357
PROBABILITY SPACES
  FAMILY OF, 40-0722
PROBLEM OF MARGINAL HOMOGENEITY, 40-0724
PROBLEM OF SYMMETRY, 41-0329
PROCESSES
  CLASS OF, 40-1856
PRODUCT DISTRIBUTION, 34-0355
PRODUCT MEASURE, 40-2218, 40-2219
PRODUCT PROBABILITY MEASURES, 41-0329
PRODUCT SPACE, 33-1480
PROOF OF ADMISSIBILITY, 40-1859
PROPERTIES OF INTEREST, 40-0722
PROPORTION OF SUCCESSES, 40-0720
PSI TEST, 34-0355
PTH ABSOLUTE CENTRAL MOMENT, 40-0721

Q

QUADRATIC MEAN, 40-1859
QUANTILE PROCESS, 40-2217
QUANTITATIVE SITUATIONS, 33-1480
QUESTION OF M-WAY MARGINAL HOMOGENEITY, 40-0724
QUEUE SIZES, 33-1502

R

RANDOM MATRICES, 34-0358
RANDOM OBSERVATION, 40-2217
RANDOM SAMPLE, 34-0355, 40-0720
RANDOM SAMPLE OF SIZE, 33-1502, 40-0723
RANDOM SUM OF EXPONENTIAL RANDOM VARIABLES, 41-0328
RANDOM VARIABLE INVARIANT IN DISTRIBUTION, 40-2216
RANDOM VARIABLES, 40-0720, 40-0722, 40-1856, 40-1859
RANDOM VECTOR, 40-0722
RANDOMNESS
  LACK OF, 33-1480
RANDOMNESS IN FINITE SEQUENCES, 33-1480
RANK ORDER, 34-0357
RANK ORDER TESTS, 40-1857
RANK ORDER TESTS STATISTICS, 40-0721
RANK ORDERS
  PROBABILITY OF, 34-0357
RANK STATISTIC, 40-1857, 41-0329
RANK TEST, 34-0355, 34-0357, 40-1857, 40-2216, 41-0329
RANK TEST OF LINEARITY VERSUS CONVEXITY, 40-1857
RANK-ORDER TEST, 41-0329

- RAPIDITY OF CONVERGENCE, 40-0723
- RATE OF CONVERGENCE, 40-0722
- REAL CASE, 40-0721
- REAL LEBESGUE A.E., 41-0329
- REAL LINE, 40-0720
- REAL STABLE MULTIVARIATE CHARACTERISTIC FUNCTION EXP, 40-1860
- REAL VARIATE VALUE, 40-2218
- REASONABLE CRITERION, 40-0722
- RECTANGULAR DISIRIBUTION, 40-0721
- RECTANGULAR VARIABLES, 34-0355
- REFLECTION PRINCIPLE, 40-1856
- REGION OF INTEREST, 40-1860
- REGRESSION CASE, 40-2217
- REGRESSION COEFFICIENT, 40-2220
- REGRESSION MODEL, 40-1857
- REGRESSION OCEFFICIENTS, 34-0356
- REGULARITY CONDITIONS, 40-2219, 41-0328, 41-0329
- RELIABILITY FUNCTIONS OF MULTI-COMPONENT STRUCTURES, 33-1480
- REPRESENTS BATCH ARRIVALS, 41-0328
- ROBERTS COEFFICIENTS, 33-1480
- ROBUST PROCEDURES, 40-1858
- ROBUST TESTS, 40-1858
- ROBUSTNESS
  - WELL-KNOWN LACK OF, 40-1858
- ROBUSTNESS OF NORMAL-THEORY TECHNIQUES, 40-1858
- ROW VECTOR, 40-2217
- ROY'S TEST, 40-1858
- RTH ABSOLUTE CENTRAL MOMENT, 40-0721

S

- SAMPLE CORRELATION COEFFICIENT, 40-0720
- SAMPLE COVARIANCE MATRIX, 40-1856
- SAMPLE EMPIRICAL PROCESS, 40-2217
- SAMPLE MEANS, 34-0356
- SAMPLE PROBLEMS, 40-2217
- SAMPLE SIZE, 40-0722, 40-1857, 40-1859, 40-2216, 41-0329
- SAMPLE SPACE, 33-1480
- SAMPLE TEST, 34-0356
- SAMPLES
  - BASIS OF, 40-2216
- SAMPLING DESIGN, 40-2218
- SAMPLING DESIGNS
  - SEQUENCE OF, 40-2218
- SAMPLING DISTRIBUTIONS
  - UNIVARIATE ONE-PARAMETER FAMILY OF, 41-0328
- SAMPLING SCHEME, 40-0722
- SATISTICAL INFERENCE ASPECT, 40-0722
- SCALAR CONSTANTS, 40-1856
- SCALE PARAMETERS, 40-0723
- SECOND-ORDER MOMENTS, 40-1858
- S PARABLE COMPLETE METRIC SPACE, 41-0330
- SEQUENCE OF ALTERNATIVES, 40-0722
- SEQUENCE OF INCLUSION PROBABILITIES, 40-2218
- SEQUENCE OF INDEPENDENT OBSERVATIONS, 40-2217
- SEQUENCE OF MEASURES, 41-0330
- SEQUENCE OF SAMPLING DESIGNS, 40-2218
- SEQUENTIAL PROCEDURE, 40-1857
- SEQUENTIAL PROCEDURES
  - CLASS OF, 40-2216
- S[illegible] RULE, 40-2216
- SEQUENTIAL SAMPLE SIZE, 40-1857
- SERIAL ASSOCIATION DIAGRAM OF, 33-1480
- SERIES EXPANSION, 40-0724
- SERVERAL TYPES OF SYMMETRIC FUNCTIONS, 40-1857
- SERVERS
  - NUMBER OF, 41-0328
- SERVICE OPERATIONS, 40-1856
- SERVICE STATIONS, 40-1856
- SERVICE TIME DISTRIBUTION, 41-0328
- SERVICE TIME DISTRIBUTION FUNCTION, 33-1502
- SERVICE TIMES, 40-1856
- SET OF DISCRETE RANDOM VARIABLES, 40-2217
- SET OF ORDER STATISTICS, 40-0723
- SET OF VALUES, 40-0720
- SETS
  - INDEPENDENCE OF, 40-1858
- SETS OF VARIATES, 40-1858
- SIMPLE CASE, 33-1480
- SIMPLE RANDOM SAMPLING, 34-0357, 40-2219
- SIMPLEST APPROACH, 33-1502
- SIMULATION TECHNIQUES, 40-0723
- SINGLE SERVER, 33-1502
- SINGLE WISHART MATRIX, 34-0358
- SIZE
  - RANDOM SAMPLE OF, 33-1502, 40-0723
- SIZES
  - INITIAL QUEUES OF, 40-1856
- SMALL SAMPLE DISTRIBUTION, 40-0723
- SMALL SAMPLE POWER, 40-1857
- SMALL SAMPLES, 40-0723
- SMALLEST INTEGER, 40-2216
- SMALLEST VARIABLE, 34-0357
- SPHERICAL CONFIDENCE REGION, 40-1857
- SQUARES
  - SUM OF, 40-0720
- STABLE PROCESSES, 40-1859
- STANDARD ERROR, 40-1856
- STATION JOIN, 40-1856
- STATIONARY INDEPENDENT INCREMENTS, 40-1856
- STATIONS LEAVE, 40-1856
- STATISTICS
  - ASYMPTOTIC THEOREMS IN, 40-2217
- STEADY STATE, 41-0328
- STOCHASTIC INTEGRAL, 40-1859
- STOCHASTIC PREDICTORS, 40-1857
- STOCHASTIC PROCESSES, 40-1859
  - CLASS OF, 40-1856
  - WEAK CONVERGENCE OF, 40-2217
- STRAIGHT LINE, 33-1502
- STRAIGHT-LINE MODEL, 40-0720
- STRAIGHTFORWARD MANNER, 34-0357
- STRUCTURE OF COVARIANCE MATRICES, 40-1858
- STUDENT'S T-STATISTIC, 40-2220
- SUBSAMPLE MEANS
  - OVERALL AVERAGE OF, 34-0357
- SUBSAMPLE SIZES, 34-0357
- SUCCESSES
  - NUMBER OF, 40-0720
  - PROPORTION OF, 40-0720
- SUCCESSIVE CLUSTERS OF EQUAL SIZE, 33-1480
- SUCCESSIVE VALUES, 33-1480
- SUFFICIENT CONDITIONS, 40-1859
- SUITABLE CONDITIONS, 40-0722
- SUITABLE GENERALIZATIONS, 40-1859
- SUITABLE TRANSOFRMATION, 34-0357
- SUM OF SQUARES, 40-0720
- SUMS OF INDEPENDENT RANDOM VARIABLES, 41-0329
- SUPREMUM FUNCTIONAL, 40-1856
- SYMMETRIC FUNCTIONS, 40-1857
  - SERVERAL TYPES OF, 40-1857
- SYMMETRIC STABLE PROCESSES
  - CHARACTERIZATIONS OF, 40-1859
- SYMMETRICAL FACTORIAL DESIGN, 40-0720, 40-0723, 40-2217, 40-2218
- SYMMETRY
  - PROBLEM OF, 41-0329
- SYSTEM IDLE TIM, 40-1856
- SYSTEMATIC SAMPLING, 34-0357

T

- TCHEBYCHEFF POLYNOMIAL, 40-1858
- TECHNICAL REPORT, 40-0723
- TELESCOPE PRINCIPLE, 40-1856
- TERRY TEST, 34-0355
- TEST CRITERIA, 40-0722
- TEST STATISTIC, 40-0722, 40-0723, 40-1858
- TESTS OF HOMOGENEITY, 40-0724
- TESTS OF HYPOTHESES, 34-0358
- TESTS OF MARGINAL HOMOGENEITY, 40-0724
- THEOREM
  - EXPECTATION IN, 41-0328
- THEOREM CONTINUES, 34-0357
- THEORY OF PROBABILITY, 34-0355
- TIME SPENT, 40-1856
- TOLERANCE REGIONS, 34-0356, 34-0358
- TOLERANCE-LIMIT PROBLEM, 34-0355
- TOPOLOGICAL ASSUMPTIONS, 33-1480
- TORONTO
  - UNIVERSITY OF, 40-2217
- TOTAL OPERATION TIME, 40-1856
- TRANSFORMATION GROUP, 40-2216
- TRANSLATION PARAMETER, 40-1859
- TRANSLATION PARAMETERS OF IDENTICAL SYMMETRIC DENSISIES, 40-1859
- TRIANGLE INEQUALITY, 40-0722
- TRIPLET OF INDICES, 34-0355
- TRUE PARAMETER, 33-1480
- TRUE PARAMETER VALUE, 40-2217
- TSCHEBY-SCHEFF'S INEQUALITY, 40-2218
- TUNNEY, 40-2220
- TWO-DIMENSIONAL RANDOM WALK REPRESENTATION, 40-1856
- TWO-SAMPLE PROBLEM, 41-0329
- TWO-SAMPLE RANK TESTS
  - POWER FUNCTIONS OF, 34-0355
- TWO-WAY CONTINGENCY TABLE, 40-0724
- TYPE II, 40-2219
- TYPES OF HISTOGRAMS, 40-1856

U

- UMP INVARIANT TEST, 40-0722
- UNCONDITIONAL DISTRIBUTION THEORY, 40-0721
- UNEQUAL SAMPLE SIZES, 34-0355
- UNIFORM PRIOR, 40-1859
- UNIFORM WEIGHTS, 40-0720
- UNIVARIATE ANALYSIS OF COVARIANCE, 40-0721
- UNIVARIATE ONE-PARAMETER FAMILY OF SAMPLING DISTRIBUTIONS, 41-0328
- UNIVERSITY OF TORONTO, 40-2217
- UPPER BOUND, 40-0723, 40-2217

V

- VALIDITY OF ANALYSIS, 34-0357
- VALUES
  - SET OF, 40-0720
- VAN DER WAERDEN-X, 41-0329
- VARIANCE ESTIMATOR, 34-0357
- VARIANCE IDENTITY
  - ANALYSIS OF, 33-1480
- VARIANCE TESTS
  - ANALYSIS OF, 34-0357
- VARIANCES
  - EQUALITY OF, 40-1858
  - HOMOGENEITY OF, 40-0722
- VARIATE VALUES, 40-2219
- VARIATES
  - NUMBER OF, 40-0721
  - SETS OF, 40-1858
- VARIATION
  - COEFFICIENTS OF, 40-0723
- VECTOR CASE, 40-0724
- VERSION OF LEHMANN'S, 34-0357
- VIA PITMAN EFFICIENCY, 40-1857
- VIEW
  - POINT OF, 41-0328

W

- WALD
  - CONDITIONS OF, 40-1857
- WEAK ASSUMPTIONS, 33-1480
- WEAK CONVERGENCE, 40-2217
  - APPLICATIONS OF, 40-2217
- WEAK CONVERGENCE OF STOCHASTIC PROCESSES, 40-2217
- WEAK SEQUENTIAL COMPACTNESS, 41-0330
- WEAKER CONDITION, 40-2217
- WEIBULL DISTRIBUTION, 33-1502
- WEIBULL PROBABILITY GRAPH PAPER, 33-1502
- WELL-KNOWN LACK OF ROBUSTNESS, 40-1858
- WELL-KNOWN SPACE, 41-0330
- WIDTH CONFIDENCE INTERVAL, 40-1859
- WIENER PROCESS, 40-1859
- WILCOXON TEST, 34-0355, 41-0329
- WILKS
  - NULL DISTRIBUTIONS OF, 40-0721

APPENDIX IV

AMALGAMATED ALGORITHMIC INDEX

ABSTRACTS IN CANCER RESEARCH

INDEX TO CANCER ABSTRACTS

A

ABLATION
  FORM OF, 65-1473
ACCESSIBLE SOFT TISSUE LESIONS, 65-1436
ACCURATE HISTOLCGIC DIAGNOSIS, 65-1409
ADDISON'S DISEASE SECCNDARY, 65-1424
ADRENAL CORTEX, 65-1462
ADRENAL INSUFFICIENCY, 65-1404
ADRENAL MAMMARY CANCER, 65-1404
ADRENAL STEROIDS, 65-1424
ADRENALECTOMY
  EFFECTS OF, 65-1404
  REVIEW OF, 65-1396
  USE OF, , 65-1410
ADRENALECTCMY APPEARS, 65-1395
ADRENALECTCMY IN CANCER,
ADRENALECTOMY MANIFESIS ITSELF, 65-1433
AGE
  YEARS OF, 65-1395, 65-1410
ALKALINE-PHOSPHATASE LEVEL, 65-1470
ALLEVIATION OF PAIN, 65-1399, 65-1423
AMPHOPHIL CELLS, 65-1424
ANTERIOR PITUITARY, 65-1462
ANTI-INFLAMMATORY ACTION, 65-1477
APPARENT EFFECT, 65-1405
APPERANCE CF PRIMARY, 65-1403
ARREST
  EVIDENCE OF, 65-1463
ARREST OF DISEASE, 65-1483
AVERAGE CURATION, 65-1395
AVERAGE DURATION OF REMISSION, 65-1463, 65-1483
AVERAGE LENGTH CF SURVIVAL, 65-1464
AVERAGE LONGEVITY, 65-1410
AVERAGE SURVIVAL, 65-1478
AVERAGE SURVIVAL TIME, 65-1476

B

BEST RESPONCE, 65-1399
BEST RESPONSES, 65-1459
BILATERAL ADRENALECTCMIES, 65-1423
BILATERAL ADRENALECTCMY, 65-1403, 65-1408, 65-1419, 65-1433
  EFFECTS OF, 65-1415
BILATERAL OOPHORECTOMY
  EFFECT OF, 65-1435
BILATERAL TOTAL ADRENALECTOMY, 65-1418
BLASTIC RESPONSE IN SKELETAL LESIONS, 65-1485
BODY HORMCNES, 65-1462
BODY WEIGHT, 65-1433
BONE LESICNS, 65-1475
BONE METASTASES, 65-1464
  RECALCIFICATION OF, 65-1436
BONE PAIN
  RELIEF IN, 65-1395
BONY METASTASES, 65-1445
BRAIN METASTASES, 65-1459
BREAST
  PALLIATION OF, 65-1424
BREAST CANCER, 65-1403, 65-1413, 65-1439, 65-1450, 65-1460, 65-1470
  CASES OF, 65-1392, 65-1419
  CHEMICAL HORMONAL CONTROL OF, 65-1484
BREAST CANCER ADRENALECTOMY, 65-1420
BREAST CANCER METASTASES, 65-1410
BREAST CANCER NOTES, 65-1396
BREAST CARCINOMA, 65-1390, 65-1461, 65-1464

C

CANCER
  ADRENALECTCMY IN,
CANCERS
  PERCENT OF, 65-1462
CARCINOMATOSIS
  CASES OF, 65-1465
CARCINOMATOUS PROSTATE
  SLIGHT REGRESSION OF, 65-1423
CASA DI CURA, 65-1392
CASE HISTCRIES, 65-1476
CASE OF HYPERTHYROIDISM, 65-1450
CASE CF PROSTATIC CARCINOMA, 65-1445
CASES CF BREAST CANCER, 65-1392, 65-1419
CASES OF CARCINOMATOSIS, 65-1465
CASTRATION IN WOMEN, 65-1450
CASTRATION THERAPY, 65-1485
CEREBRAL METASTASES, 65-1390
CESSATION OF HEMOPTYSIS, 65-1390
CESSATION OF PAIN, 65-1390
CHEMICAL FORMONAL CONTROL OF BREAST CANCER, 65-1484
CHIEF SIDE EFFECTS, 65-1463
CHIEF TOXIC EFFECT, 65-1479
CHOROIDAL METASTASIS, 65-1443
CLINICAL COURSE, 65-1435
CLINICAL STATE, 65-1435, 65-1476
CCMBINATICN CF DIAGNOSTIC THORACOTCMY, 65-1409
COMPARATIVE STUDY, 65-1452
COMPLETE ALLEVIATION OF PAIN, 65-1458
COMPLETE PAIN RELIEF, 65-1461
COMPLETE REGRESSION, 65-1418
COMPLICATION OF HORMONE IMBALANCE, 65-1450
CONCCMITANT TRANSDIAPHRAGMATIC UNILATERAL ADRENALECTOMY, 65-1409
CONSIDERABLE IMPROVEMENT, 65-1418
CONTRLL GROUP, 65-1452
CONTROL OF DIABETES INSIPIDUS, 65-1485
COURSE OF DISEASE, 65-1405
COURSE OF ESTROGENS PLUS RADICTHERAPY, 65-1445
CUMULATIVE LONGEVITY, 65-1410

D

DAY ICTERUS, 65-1473
DEFINITE BENEFICIAL EFFECT, 65-1419
DEFINITE CLINICAL IMPROVEMENT, 65-1465
DEGREE OF PALLIATION, 65-1404
DEGREE OF THYROIDAL DEPRESSION, 65-1485
DEGREES OF OBJECTIVE REGRESSION, 65-1423
DEMONSTRABLE TUMOR, 65-1418
DESIRABLE TREATMENT, 65-1479
DESOXYCORTICOSTERONE
  DOSES OF, 65-1418
DESTRUCTION OF RESIDUAL ANTERIOR LOBE TISSUE, 65-1463
DEVELOPMENT OF MAMMARY CARCINOMA DNO, 65-1436
DEVELOPMENT OF MENINGITIS, 65-1465
DIABETES INSIPIDUS, 65-1463, 65-1479
  CONTROL OF, 65-1485
DIAGNOSTIC THORACOTOMY
  COMBINATION OF, 65-1409
DIMUTION IN SIZE, 65-1390
DISAPPEARANCE OF LOCAL, 65-1461
DISAPPEARANCE OF PLEURAL EFFUSIONS, 65-1390
DISAPPEARANCE OF SOFT TISSUE LESIONS, 65-1436
DISCONTINUATION OF TREATMENT, 65-1445
DISEASE
  ARREST OF, 65-1483
  COURSE OF, 65-1405
DISTANT METASTASES, 65-1461, 65-1484
DOSES OF DESOXYCORTICOSTERONE, 65-1418
DRAMATIC REGRESSIVE, 65-1423
DROP IN ESTROGEN, 65-1407
DURATION OF REMISSIONS, 65-1436
DURATION OF SYMPTONS, 65-

E

EFFECT OF BILATERAL OOPHORECTOMY, 65-1435
EFFECT OF HYPOPHYSECTOMY, 65-1464
EFFECTIVE PALLIATIVE TREATMENT, 65-1392
EFFECTS OF ADRENALECTOMY, 65-1404
EFFECTS OF BILATERAL ADRENALECTCMY, 65-1415
EFFECTS OF HYPOPHYSECTCMY, 65-1463, 65-1476
EFFECTS CF SURGICAL, 65-1439
EIGHTEEN PATIENTS, 65-1410
ELECTROLYTE HEMEOSTASIS, 65-1477
ENDOCRINE ABLATIONS, 65-1405
ENDOCRINE CONTROL
  METHODS OF, 65-1467
ENDOCRINE DEPENDENCE OF TUMORS, 65-1467
ENDOCRINE GLANDS
  REMOVAL OF, 65-1462
ENDOCRINE GLANDS OF SIX PATIENTS, 65-1424
ENDOCRINE PATTERN, 65-1424
ENDOCRINE THERAPY, 65-1478
ENDOCRINE THERAPY OFFERS
  FORM OF, 65-1399
ESTROGEN
  DROP IN, 65-1407
ESTROGEN ADMINISTRATION, 65-1410
ESTROGEN DEPRIVATION, 65-1436
ESTROGEN EXCRETION, 65-1435, 65-1476
ESTROGEN EXCRETION FELL, 65-1435
ESTROGEN LEVEL
  REDUCTION IN, 65-1435
ESTROGEN LEVELS, 65-1435
ESTROGEN SECRETION, 65-1435
ESTROGEN THERAPY, 65-1423, 65-1445
ESTROGENIC SUBSTANCES
  IMPORTANCE OF, 65-1436
ESTROGENS PLUS RADIOTHERAPY
  COURSE OF, 65-1445
ETIOLOGICAL FACTOR, 65-1473
EVALUABLE INITIAL OBJECTIVE REMISSIONS, 65-1484
EVALUATION OF PALLIATIVE TREATMENT, 65-1408
EVIDENCE OF ARREST, 65-1463
EVIDENCE OF REACTIVATION, 65-1436
EVIDENCE OF SPINAL REGENERATION, 65-1470
EXCEPTION OF HYPOPHYSECTOMY, 65-1399
EXCRETE ESTROGEN, 65-1476

F

- FALL IN URINARY CALCIUM, 65-1485
- FAVORABLE RESPONSE, 65-1436
- FIFTY PERCENT, 65-1463
- FIVE PATIENTS, 65-1424
- FORM OF ABLATION, 65-1473
- FORM OF ENDOCRINE THERAPY OFFERS, 65-1395
- FORM OF THERAPY, 65-1436
- FOUR OF SIX PATIENTS, 65-1435
- FOUR OF SUCH, 65-1462
- FOUR PATIENTS, 65-1435, 65-1483
- FOUR WOMEN, 65-1476

G

- GAIN IN WEIGHT, 65-1458, 65-1479
- GAIN OF WEIGHT, 65-1485
- GALLIERA SARCOMA, 65-1433
- GLAND PARENCHYMA, 65-1463
- GRAVES' DISEASE, 65-1450
- GREW WORSE, 65-1403
- GROUPS OF PATIENTS, 65-1452
- GROWTH HORMONE, 65-1473

H

- HEMOLYSIS
  - OCCURRENCE OF, 65-1473
- HEMOPTYSIS
  - CESSATION OF, 65-1390
- HEPATIC INVOLVEMENT, 65-1422
- HEPATIC METASTASIS
  - CONTRAINDICATES ADRENALECTOMY, 65-1422
- HIGH PRE-OPERATIVE ESTROGEN EXCRETION, 65-1476
- HIGH PRIOR, 65-1423
- HISTOLOGICAL STUDY, 65-1424
- HISTOLOGY OF TUMOR, 65-1408
- HISTORICAL INTRODUCTION,
- HORMONAL REPLACMENT THERAPY, 65-1404
- HORMONE IMBALANCE
  - COMPLICATION OF, 65-1450
- HORMONE SUPPRESSION, 65-1450
- HORMONE THERAPY, 65-1452, 65-1462
- HORMONES
  - INFLUENCE OF, 65-1395
- HUMAN TUMORS, 65-1424
- HUMANS
  - PROSTATE IN, 65-1462
- HYPERTHYROIDISM
  - CASE OF, 65-1450
- HYPOPHYSEAL STALK, 65-1483
- HYPOPHYSECTOMY
  - EFFECT OF, 65-1464
  - EFFECTS OF, 65-1463, 65-1476
  - EXCEPTION OF, 65-1399
  - VALUE OF, 65-1460

I

- IMMEDIATE PAIN RELIEF, 65-1470
- IMMEDIATE POSTOPERATIVE PERIOD, 65-1436
- IMPORTANCE OF ESTROGENIC SUBSTANCES, 65-1436
- IMPROVEMENT
  - OBJECTIVE EVIDENCE OF, 65-1478
- IMPROVEMENT OF VISUAL FIELDS, 65-1390
- INCREASES APPETITE, 65-1390
- INDICATIONS OF SUBJECTIVE RESPONSE, 65-1390
- INDIVIDUAL PROBLEM, 65-1418
- INFLUENCE OF HORMONES, 65-1395
- INHIBITION OF TUMOR GROWTH PREFACES, 65-1395
- INTERSTITIAL IRRADIATION, 65-1396
- INTRATHORACIC LESIONS, 65-
- IRRADIATION HYPOPHYSECTOMY, 65-1461

K

- KAROLINSKA HOSPITAL IN STOCKHOLM, 65-1392

L

- LATTER PROCEDURE, 65-1462
- LENGTHS OF TIME, 65-1461
- LESION
  - SIZE OF, 65-1452
- LOCAL
  - DISAPPEARANCE OF, 65-1461
- LOCAL DISEASE, 65-1484
- LOCAL PROSTATIC CANCER, 65-1475
- LONGEST SINGLE SURVIVAL, 65-1390
- LOW ADRENOCORTICAL ACTIVITY, 65-1407
- LOW PRE-OPERATIVE LEVELS OF URINARY ESTROGEN, 65-1476

M

- MAINTENANCE DOSE, 65-1485
- MAINTENANCE THERAPY, 65-1418, 65-1479
- MAJOR BENEFIT, 65-1470
- MALE PATIENT, 65-1476
- MALIGNANCIES
  - TREATMENT OF, 65-1401
- MALIGNANT TUMORS, 65-1454
- MAMMARY CANCER, 65-1442
- MAMMARY CARCINOMA, 65-1407, 65-1452
- MAMMARY CARCINOMA AND
  - DEVELOPMENT OF, 65-1436
- MASSES
  - SIZE OF, 65-1390
- MASTECTOMY
  - TIME OF, 65-1452
- MEAN DURATION, 65-1404
- MEAN SURVIVAL OF PATIENTS, 65-1462
- MEAN SURVIVAL TIME, 65-1390, 65-1419
- MENINGITIS
  - DEVELOPMENT OF, 65-1465
- MENOPAUSAL GROUP, 65-1435
- MENOPAUSAL PATIENTS, 65-1459
- METASTATIC BREAST CANCER, 65-1390, 65-1405, 65-1418, 65-1422, 65-1435, 65-1458, 65-1476
- METASTATIC BREAST CARCINOMA, 65-1463, 65-1477
- METASTATIC CANCER, 65-1419
- METASTATIC CARCINOMA, , 65-1424
- METASTATIC DEPOSITS
  - REGENERATION IN, 65-1470
- METASTATIC GROWTH
  - REGRESSION OF, 65-1459
- METASTATIC INHIBITION, 65-1458
- METASTATIC LESIONS, 65-1390, 65-1445
- METASTATIC MAMMARY CARCINOMA, 65-1483, 65-1485
- METHODS OF ENDOCRINE CONTROL, 65-1467
- MONTH PERIOD, 65-1390
- MONTHS IN RELATIVE COMFORT, 65-1471

N

- NEOPLASM
  - REGRESSION OF, 65-1485
- NODE INVOLVEMENT, 65-1452
- NORMAL ACTIVITY, 65-1422, 65-1485
- NORMAL LEVELS IN PATIENTS, 65-1424
- NORMAL LIFE, 65-1465
- NORMALIZATION OF TEMPERATURE, 65-1459
- NOTEWORTHY IMPROVEMENT, 65-1464
- NUCLEAR PYKNOSIS, 65-1424
- NUMBER OF REMISSIONS, 65-1419

O

- OBJECTIVE BENEFITS, 65-1475
- OBJECTIVE EVIDENCE, 65-1390
- OBJECTIVE EVIDENCE OF IMPROVEMENT, 65-1478
- OBJECTIVE IMPROVEMENT, 65-1390, 65-1436
- OBJECTIVE REGRESSION
  - DEGREES OF, 65-1423
- OBJECTIVE REGRESSIONS, 65-1435
- OBJECTIVE REMISSION, 65-1395, 65-1396, 65-1408, 65-1435, 65-1471
- OBJECTIVE RESPONSE, 65-1422
- OBJECTIVE SIGNS OF REGRESSION, 65-1475
- OCCURRENCE OF HEMOLYSIS, 65-1473
- OESTROGEN PRODUCTION, 65-1435
- OOPHORECTOMY
  - OVERALL EFFECT OF, 65-1435
- OOPHORECTOMY IN POSTMENOPAUSAL WOMEN, 65-1435
- OPERATIVE MORTALITY, 65-1390, 65-1465
- OPERATIVE TECHNIQUE, 65-1390, 65-1465
- OPPOSITE ADRENAL GLAND, 65-1409
- OSSEOUS LESIONS
  - RAPID REPAIR OF, 65-1454
- OSTEOLYTIC LESIONS
  - SCLEROSIS OF, 65-1390
- OVARIAN CASTRATION
  - THERAPEUTIC POSSIBILITIES OF, 65-1438
- OVERALL EFFECT OF OOPHORECTOMY, 65-1435

P

- PAIN
  - ALLEVIATION OF 65-1399, 65-1423
  - CESSATION OF, 65-1390
  - COMPLETE ALLEVIATION OF, 65-1458
  - PARTIAL RELIEF OF, 65-1403
  - RELIEF OF, 65-1418, 65-1475
- PALLIATION
  - DEGREE OF, 65-1404
  - PROSPECT OF, 65-1399
- PALLIATION OF BREAST, 65-1424
- PALLIATIVE BENEFITS, 65-1461
- PALLIATIVE TREATMENT, 65-1479
  - EVALUATION OF, 65-1408
- PARTIAL RELIEF OF PAIN, 65-1403
- PATIENT FOUR YEARS, 65-1443
- PATIENTS
  - GROUPS OF, 65-1452
  - MEAN SURVIVAL OF, 65-1462
  - NORMAL LEVELS IN, 65-1424
  - PERCENTAGE OF, 65-1392
  - SELECTION OF, 65-1408
  - SERIES OF, 65-1413
  - VISUAL FIELDS IN, 65-1390
- PATIENTS ALIVE, 65-1462
- PATIENTS ESTROGEN EXCRETION, 65-1476
- PATIENTS UNFIT, 65-1396
- PELVIC BONE LESIONS, 65-1403
- PELVIC COMPLICATIONS NODE INVOLVEMENT, 65-1452
- PELVIC REGION, 65-1454
- PERCENT OF CANCERS, 65-1462
- PERCENT OF THREE-MONTH SURVIVORS, 65-1462
- PERCENTAGE OF PATIENTS, 65-1392
- PERIODS OF REMISSION, 65-1459
- PERITONEAL METASTASES, 65-1464
- PITUITARY STALK, 65-1479
- PITUITARY STALK SECTION, 65-1485
- PLEURAL EFFUSIONS
  - DISAPPEARANCE OF, 65-1390
- PLEURAL METASTASES, 65-1464
- POST OPERATIVE COMPLICATIONS, 65-1404
- POST- MENOPAUSE, 65-1408
- POST-OPERATIVE BIOCHEMICAL CONTROL, 65-1461
- POST-OPERATIVE PERIOD, 65-1464
- POST-OPERATIVE X-RAYS, 65-1470
- POSTMENOPAUSAL WOMEN
  - OOPHORECTOMY IN, 65-1435
- POSTMORTEM STUDIES, 65-1463
- POSTOPERATIVE CARE, 65-1390
- POSTOPERATIVE DEATH, 65-1419
- POSTOPERATIVE MANAGEMENT, 65-1408
- POSTOPERATIVE SUBSTITUTION THERAPY,
- PRE
  - STATE OF, 65-1408
- PRE-MENOPAUSAL CASES, 65-1435
- PRELIMINARY REPORT, 65-1477
- PREMENOPAUSAL PATIENT, 65-1419
- PREOPERATIVE BIOCHEMICAL ASSESSMENT, 65-1461
- PRIMARY
  - APPERANCE OF, 65-1403
- PRIMARY ADDISON'S DISEASE, 65-1424
- PRIMARY ADRENAL INSUFFICIENCY, 65-1424
- PRIMARY GROWTH, 65-1399
- PRIMARY HYPOPHYSECTOMY, 65-1410
- PRIMARY LESION, 65-1470
- PROBLEM OF REHABILITATION, 65-1418
- PRODUCE ACTH THYROTROPIN, 65-1424
- PRODUCE CLINICAL BENEFIT, 65-1462
- PRODUCE COMPENSATORY HYPOPHYSEAL OVERACTIVITY, 65-1424
- PROGRESSIVE METASTATIC BREAST CANCER, 65-1395, 65-1408
- PROLONGATION OF SURVIVAL, 65-1431
- PROMPT RELIEF, 65-1470
- PROSPECT OF PALLIATION, 65-1399
- PROSTATE CANCER, 65-1475
- PROSTATE GLAND, 65-1431
- PROSTATE IN HUMANS, 65-1462
- PROSTATE TUMOR, 65-1445
- PROSTATIC CANCER, 65-1413, 65-1415, 65-1462
- PROSTATIC CARCINOMA
  - CASE OF, 65-1445
- PROSTATIC CARCINOMAS, 65-1424

R

- RADICAL MASTECTOMY AG, 65-1452
- RADIOACTIVE GOLD, 65-1396
- RADIOACTIVE GOLD SEEDS, 65-1463
- RADIOACTIVE YTTRIUM, 65-1465
- RAPID REPAIR OF OSSEOUS LESIONS, 65-1454
- RAPID SKELETAL INVOLVEMENT, 65-1405
- RAPID TREATMENT, 65-1409
- REACTIVATION
  - EVIDENCE OF, 65-1436
- RECALCIFICATION OF BONE METASTASES, 65-1430
- RECURRENT BREAST CANCER, 65-1464

- RECURRENT METASTATIC CARCINOMA, 65-1409
- REDUCTION IN ESTROGEN LEVEL, 65-1435
- REDUCTION OF SURGICAL TRAUMA, 65-1409
- REGENERATION IN METASTATIC DEPOSITS, 65-1470
- REGIONAL LYMPH NODES BECAME IMPALPABLE, 65-1470
- REGRESSION
  - OBJECTIVE SIGNS OF, 65-1475
- REGRESSION OF METASTATIC GROWTH, 65-1459
- REGRESSION OF NEOPLASM, 65-1489
- REGRESSION OF SKIN METASTASES, 65-1458
- REGRESSION OF TUMOR, 65-1423
- REHABILITATION
  - PROBLEM OF, 65-1418
- RELATIVE COMFORT
  - MONTHS IN, 65-1471
- RELIEF IN BONE PAIN, 65-1395
- RELIEF OF PAIN, 65-1418, 65-1475
- REMISSION
  - AVERAGE DURATION OF, 65-1403, 65-1433
  - PERIODS OF, 65-1459
- REMISSIONS
  - DURATION OF, 65-1436
  - NUMBER OF, 65-1415
- REMOVAL OF ENDOCRINE GLANDS, 65-1452
- RESIDUAL ANTERIOR LOBE TISSUE
  - DESTRUCTION OF, 65-1463
- SPONSIVE NEOPLASMS, 65-1410
- TINAL DETACHMENT, 65-1443
- EVIEW OF ADRENALECTOMY, 65-1396
- ISK APPEARS, 65-1465
- ENTGEN THERAPY, 65-1443
- LE OF THYROID THERAPY, 65-1439

S

- SCLEROSIS OF OSTEOLYTIC LESIONS, 65-1390
- SECONDARY GROWTHS, 65-1403
- SELECTION OF PATIENTS, 65-1408
- SERIES OF PATIENTS, 65-1413
- SERUM ACID PHOSPHATASE, 65-1423
- SEVEN OPERATIVE DEATHS, 65-1464
- SEVENTY-NINE PATIENTS, 65-1390
- SEVERE GASTRIC DISTURBANCES, 65-1471
- SHORT DURATION, 65-1479
- SHORT SURVIVAL, 65-1475
- SHORT TIME, 65-1478
- SIDE EFFECTS, 65-1471
- SIMPLE ENDOCRINE PROCEDURES, 65-1402
- SIX CASES, 65-1470
- SIX PATIENTS, 65-1403, 65-1405, 65-1483
  - ENDOCRINE GLANDS OF, 65-1424
  - FOUR OF, 65-1465
- SIX POST OPERATIVE MORTALITIES, 65-1403
- SIZE
  - DIMUTION IN, 65-1390
- SIZE OF LESION, 65-1452
- SIZE OF MASSES, 65-1390
- SKELETAL LESIONS
  - BLASTIC RESPONSE IN, 65-1405
- SKELETAL METASTASES, 65-1399
- SKIN METASTASES
  - REGRESSION OF, 65-1458
- SLIGHT REGRESSION OF CARCINOMATOUS PROSTATE, 65-1423
- ETAL REPLACEMENT,
- SOFT TISSUE INVASION, 65-1405
- SOFT TISSUE LESIONS, 65-1404
  - DISAPPEARANCE OF, 65-1436
- SPINAL CORD METASTASES, 65-1418
- SPINAL REGENERATION
  - EVIDENCE OF, 65-1470
- SPLEEN-ADRENAL VENOVENOSTOMY, 65-1407
- STATE OF FREE, 65-1408
- STEROID HORMONES, 65-1424
- STEROID METABOLISM, 65-1410
- STILBESTROL TREATMENT, 65-1462
- STOCKHOLM
  - KAROLINSKA HOSPITAL IN, 65-1392
- SUBJECTIVE BENEFITS, 65-1404
- SUBJECTIVE IMPROVEMENT, 65-1390, 65-1395, 65-1396, 65-1401, 65-1422, 65-1454, 65-1464, 65-1471, 65-1485
- SUBJECTIVE RESPONSE, 65-1436
  - INDICATIONS OF, 65-1390
- SUBSEQUENT WEIGHT GAIN, 65-1470
- SUCCESSIVE ATTEMPTS, 65-1410
- SURGICAL
  - EFFECTS OF, 65-1439
- SURGICAL CASTRATION, 65-1436, 65-1445, 65-1452
- SURGICAL MORTALITY, 65-1408
- SURGICAL TRAUMA
  - REDUCTION OF, 65-1409
- SURICAL TECHNIQUES,
- SURVIVAL
  - AVERAGE LENGTH OF, 65-1464
  - PROLONGATION OF, 65-1431
- SURVIVAL RATES, 65-1452
- SURVIVAL RATES EXCEPT, 65-1452
- SYMPTONS
  - DURATION OF, 65-1452

T

- TABULAR FORM, 65-1419
- TEMPERATURE
  - NORMALIZATION OF, 65-1459
- TEMPORARY DISTURBANCES OF VISION, 65-1471
- TEMPORARY REMISSION, 65-1424
- TEN PATIENTS, 65-1422, 65-1476
- TESTOSTERONE TREATMENT, 65-1408
- THEORETICAL CONSIDERATIONS, 65-1436
- THERAPEUTIC POSSIBILITIES OF OVARIAN CASTRATION, 65-1438
- THERAPEUTIC PURPOSES, 65-1422
- THERAPEUTIC VALUE, 65-1395
- THERAPY
  - FORM OF, 65-1438
- THEREFORE ADRENALECTOMY, 65-1392
- THIRD PATIENT, 65-1418, 65-1470, 65-1475
- THIRTY-THREE PATIENTS, 65-1464
- THOUGHT WORTH, 65-1465
- THREE-MONTH SURVIVORS
  - PERCENT OF, 65-1462
- THYROID EXTRACT, 65-1464
- THYROID THERAPY
  - ROLE OF, 65-1439
- THYROIDAL DEPRESSION
  - DEGREE OF, 65-1465
- TIME
  - LENGTHS OF, 65-1461
- TIME OF MASTECTOMY, 65-1452
- TOTAL HYPOPHYSECTOMY, 65-1479
- TRANSIENT DIABETES, 65-1471
- TRANSIENT SUBJECTIVE IMPROVEMENT, 65-1422
- TREATMENT
  - DISCONTINUATION OF, 65-1445
- TREATMENT OF MALIGNANCIES, 65-1461
- TREATMENT PREFACE, 65-1436
- TUMOR
  - HISTOLOGY OF, 65-1408
  - REGRESSION OF, 65-1423
- TUMOR CELLS, 65-1433
- TUMOR DEVELOPMENT, 65-1433
- TUMOR GROWTH, 65-1401, 65-1404
- TUMOR GROWTH PREFACES
  - INHIBITION OF, 65-1395
- TUMOR ITSELF, 65-1473
- TUMOR SIZE, 65-1407
- TUMORS
  - ENDOCRINE DEPENDENCE OF, 65-1467

U

- UNDERGONE ADRENALECTOMIES, 65-1413, 65-1424
- UNILATERAL ADRENALECTOMY, 65-1407
- UNINARY ESTROGEN
  - LOW PRE-OPERATIVE LEVELS OF, 65-1476
- URINARY CALCIUM
  - FALL IN, 65-1485
- USE OF ADRENALECTOMY, ; 65-1419

V

- VALUE OF HYPOPHYSECTOMY, 65-1460
- VENOUS HEMORRHAGE, 65-1404
- VISCERAL METASTASES, 65-1399
- VISION
  - TEMPORARY DISTURBANCES OF, 65-1471
- VISUAL FIELDS
  - IMPROVEMENT OF, 65-1390
- VISUAL FIELDS IN PATIENTS, 65-1390

W

- WEEKS LATER, 65-1409
- WEIGHT
  - GAIN IN, 65-1458, 65-1475
  - GAIN OF, 65-1485
- WEIGHT GAIN, 65-1395, 65-1403
- WHETHER HYPOPHYSECTOMY, 65-1461
- WHOM
  - FOUR OF, 65-1462
- WIDESPREAD METASTASES, 65-1473, 65-1476
- WIDESPREAD METASTATIC BREAST CANCER, 65-1462
- WOMEN
  - CASTRATION IN, 65-1450
- WORTHWHILE PALLIATION, 65-1420

Y

- YEAR LATER, 65-1409
- YEARS LATER, 65-1410
- YEARS OF AGE, 65-1395, 65-1410